Published in final edited form as:
J Appl Res Mem Cogn. 2023 June ; 12(2): 199–202. doi:10.1037/mac0000125.

Recognizing limits on the generalizability of findings of psychological science research

Patricia J. Bauer

Department of Psychology, Emory University

Abstract

The five commentaries on the target article “Generalizations: The Grail and the Gremlins” (Bauer, 2023) provide food for thought on the issue of generalizations in psychological science. Generally speaking, there seems to be agreement in the field that unrecognized limits on the generalizability of the findings of psychological science research are a matter of serious concern. This reply to the commentaries features elaboration of this basic point of agreement as well as discussion of other points of convergence with the arguments in the target article. The reply also addresses areas of divergence between the target article and the commentaries, and among the commentaries themselves. The reply develops suggestions for ways forward inspired by the commentaries. Echoing the target article, the reply calls for greater authenticity in psychological science research, movement toward which would strengthen efforts along the full range of inquiry, from basic to applied.

Keywords

applied research; ecological validity; generalizability; individual variability; research practices; sampling

My thanks to the five authors and author groups who provided commentaries on the target article “Generalizations: The Grail and the Gremlins” (Bauer, 2023). I appreciate the thought the authors put into their observations, critiques, and suggestions. It is clear from the comments that generally speaking, there is agreement in the field that potential unrecognized limits on the generalizability of the findings of psychological science are a matter of serious concern. In what follows, I elaborate on this and other points of convergence between the target article and the commentaries. I then take up points of divergence between the target article and the commentaries, and among the commentaries themselves. Finally, I develop some suggestions for ways forward.

Areas of Convergence

Across the commentaries, there is substantial convergence with the arguments in the target article (as well as in Yarkoni, 2022) that generalizability is a hallmark of psychological science and that at present, the field falls short of achieving this goal. The authors of the

Correspondence may be addressed to Patricia J. Bauer, Department of Psychology, 36 Eagle Row, Emory University, Atlanta, Georgia 30322, USA; pjbauer@emory.edu.

commentaries also highlight that these issues are especially critical when we have in mind not just behavior in the laboratory, but in the world beyond it—that is, as we attempt to apply the findings of our research to classrooms, workplaces, and courtrooms, to name a few (Mitchell & Shivde, 2023; Otgaar et al., 2023). They reinforce the need for attention to cultural background, individual variability, methodological variations, and other sources of variance between and among participants that may limit generalizability of findings (Peterson, 2023; Prather, 2023). In short, there is general agreement that it is timely and important to attend to issues of generalizability (Devezer & Buzbas, 2023).

The commentary authors also make important points not taken up in the target article, but that are consistent with its message. One such observation is that as a field, we do not do a very good job of distinguishing between statistical significance and practical (or I will add, theoretical) relevance. Otgaar and colleagues (2023) make this point in the realm of eyewitness testimony and forensic applications more broadly. They note that sometimes in forensic applications, small differences, such as a single forgotten or incorrect detail, can have profound implications for evaluation of the veracity of eyewitness identifications or the outcome of legal proceedings, for example. They advocate for use of analyses that turn on the "smallest effect size of interest" (e.g., Anvari & Lakens, 2021) as opposed to the more often invoked Cohen's *d* or other index of the size of an effect. Both in the target article and here, I suggest the complement as well—that we not draw strong conclusions about group or condition differences that are themselves small; that may be revealed only with the support of exceptionally well-powered research designs; and which may apply to the groups, but mischaracterize the individuals in the groups. As Otgaar et al. point out, statistical significance is not the same as practical relevance. Conclusions based on mean differences, or lack thereof, have consequences, and we need to be circumspect as we draw them.

Another point that was not taken up in the target article, but that is consistent with its general message, was noted by Mitchell and Shivde (2023), namely, that as a field, we seem to suffer from hesitation at discussing the limitations of our work broadly, and potential limits on generalizability specifically. We seem to be hesitant, even though we know it is the right thing to do. As Mitchell and Shivde note, training in psychological science drills new researchers in evaluation of the strengths and weaknesses of our own and others' research, and the need to discuss those limitations. Yet somewhere along the way, we lose sight of that necessity. As we work to establish ourselves as productive scholars and researchers, advance our reputations, and demonstrate that our work is innovative and impactful, we seemingly shrink from admitting that each experiment or study we conduct and publish suffers from limitations, some of which are on generalizability of findings. Using the Müller-Lyer illusion as an example (Müller-Lyer, 1889), Prather (2023) explicitly reminds us that findings may be interesting and important, even if they are not universal. And when they are not universal —or may not be—it is incumbent upon us to say so.

Mitchell and Shivde (2023) also make the important point that issues of generalizability may vary at different levels of analysis or at different points in a research program. If the goal of a research project is to establish new ground by investigating a new phenomenon, then tests of whether the findings from the initial experiments generalize to other samples, populations, and methods may be premature. Yet we should not grow complacent, lest we lapse into advancing a “truth” about behavior that falls short of that status. Peterson (2023) provides an example of this in the domain of childhood amnesia: the relative paucity among adults (and older children) of memories of specific events from the first 3 to 4 years of their lives. When this observation was first made in the late 19th century (Henri & Henri, 1895, 1898; Miles, 1893), it was a great place to start to investigate this interesting aspect of the distribution of memories across the lifespan. Yet as Peterson points out, although there now has been over a century of research on the age of earliest memory among both adults and children, we still cling to this simplification, even though some findings now shake the foundation of this timeframe. The net effect is that both our applications of the work and our theorizing about the phenomenon itself suffer.

Areas of Divergence

The authors of the commentaries are in general agreement with the main message of the target article, which is that unrecognized limits on the generalizability of the findings of psychological science are a matter of serious concern. The commentaries also feature points of divergence between the target article and the commentaries, and among the commentaries themselves. Some of the seemingly divergent views between the target article and the commentaries center on characterizations of the arguments in the target article that simply are not warranted. Because these characterizations may present roadblocks to consensus on ways forward to address issues of generalizability, it is important to discuss them. After doing so, I move on to other points of divergence in perspectives.

The first unwarranted characterization is that in the target article, the problem of generalization is framed as separate from issues of rigor and replicability (Devezer & Buzbas, 2023). Contrary to Devezer and Buzbas’ assertion, I did not argue that generalization can be achieved “independently from statistical validity” (p. 3, **TYPESETTER CHANGE TO ACTUAL PAGE NUMBER**). Obviously, the two ideals must go hand-in-hand. I did not address issues of rigor and replicability—or take up arguments regarding Null Hypothesis Statistical Testing—for the reason stated in the target article, namely, that there are excellent treatments of these issues already in the literature (e.g., Eich, 2014; Fife, 2020; Lindsay, 2015; Nosek et al., 2015; Simmons et al., 2011). Rather than attempt to further those treatments, I elected to work to make a different point—that the most rigorous and replicable findings may nevertheless be of limited value if they fail to inform how people live their lives and address true challenges and problems. In other words, if our work is not authentic, it really does not matter whether it is rigorous and can be reproduced. I stand by this perspective, especially in the domain of “applied” research.

The second unwarranted characterization is that the target article argues the case that “studies should produce easy to generalize results,” based on a “normative” human who is an able-bodied white man, and that failures to generalize beyond this “generic default person” are interpreted as evidence of deficits in the populations who do not look like, sound like, or behave like the “default” or “normative” human (Prather, 2023, p. 4 **TYPESETTER CHANGE TO ACTUAL PAGE NUMBER**). Contrary to this assertion, the target article does not suggest that the goal of psychological science is to search for universals or identify normative behavior (the terms “universals” and “normative” are not used in the target article). Rather, the perspective championed in the target article is that a major goal of psychological science is to determine the *limits* on generalizability. That is, we need to determine to whom our results apply, when, and under what circumstances. This message is precisely the opposite that Prather’s commentary attributed to the target article.

Having “defended” the target article against unwarranted charges against it, I turn to material differences of opinion that could stand in the way of progress to address concerns over generalizability. The most prominent of these differences emerges along the lines of the proverbial argument regarding which goals should dominate in psychological science: those of application or those of theory. Otgaar and colleagues (2023) make the argument that the practical relevance of psychological research is paramount. Indeed, they criticize the target article for not paying sufficient attention to this issue in the discussion of generalizability. Conversely, Devezer and Buzbas (2023) argue for the value of a “model-centric” approach. They criticize the target article for advancing the perspective that psychological science is merely a collection of facts not bound by theory, and for not paying sufficient attention to the role of theory in the scientific process. The “glass is half full” way of looking at the divergence of these perspectives is that the target article struck just the right balance—advocating for neither of the extreme positions that characterize our field.

For the record, practical relevance or significance certainly is an important consideration, especially as we evaluate research intended for application or which can reasonably be expected to be applied beyond the walls of the laboratory. Yet as discussed by Sternberg and Gordeeva (1996), scientific findings have many ways of being impactful, only one of which is through their practical application—another is through their theoretical significance (other important factors identified in the study were the quality of presentation of the findings, both substantive and methodological interest, and heuristic value). Regardless of their other attributes, to be impactful, scientific findings must be understood in terms of potential limits on their generalizability. For the flip-side of the record, a model-centric approach may be an ideal, but not all scientific progress moves from theory to observation. Especially when phenomena are young, there may not be a coherent theory or model to guide inquiry. Instead, a body of observations may need to be collected before coherence can be brought to them in the form of a model that then can guide the next steps of the process. Devezer and Buzbas (2023) are right that psychological science is more than a collection of facts. But sometimes it can—and must—start in precisely that way.

Charting a Way Forward

Having agreed that issues of generalizability are serious in psychological science and on some of their sources, it now is time to look to the future—both immediate and longer-term—to figure out how to address them. The commentaries feature and/or inspire several fruitful ways forward.

The first suggestion for a way forward is for a course of action that is both obvious and attainable with a modest shift of effort, namely, that we as researchers do more to acknowledge the potential limitations of our work in terms of its generalizability beyond the specific populations sampled, the stimuli and procedures used, the analyses applied, and so forth. Even in journals that govern the number of words that can be devoted to introductory and discussion material (e.g., *Psychological Science*), we can and should make room for these reflections. Discussions of potential limitations on the generalizability of our findings are just as important as discussions of the strengths of our efforts; discussions of potential limits on generalizability may have as much if not more heuristic value for the field. As noted by Mitchell and Shivde (2023), we were taught to pay attention to and "own" the limitations of our work. We need to devote more effort to not just learning, but to living that lesson.

A second obvious, attainable, and desirable change in behavior is to engage in more explicit evaluation of the theorical and practical implications of our findings; implications need not relate to issues of generalizability, but they often do. In evaluating the importance of their work, we tell our students not to use small *p* values as the foundation for claims that our findings are "highly significant." We preach that instead, they should look to the size of the group difference, the amount of variance explained, and so forth. Yet all too often, when we talk about our own findings, we fail to apply these very metrics. We can and should take the size of effect into consideration especially as we move from the laboratory to the classroom or courtroom. And as Otgaar and colleagues (2023) remind us, the effect need not be large to be of consequence (though small effects may be especially likely to fail tests for generalization). Effect sizes also have theoretical implications. As Peterson (2023) notes, some methodological variations in elicitation of age of earliest memory have pronounced effects that must be taken into consideration as we generalize about age of earliest autobiographical memory, and consider the theories that account for their relative paucity in the early years of life. These actions require relatively small changes in behavior that could have large positive consequences.

Third, and related to the point about consideration of the implications of our findings, is that we need to be more cognizant of the meanings or implications of our findings in different contexts. Even when effects generalize across populations and stimuli, for example, they may mean something different in different contexts. Otgaar and colleagues (2023) made this point regarding applications of laboratory findings to forensic contexts. We also may think about different implications of the same finding to different populations. Consider that an intervention may have the same positive effect in samples drawn from two different populations and in that regard, the effect generalizes. However, the implications of the positive change may be quite different for a group already high in the attribute or behavior relative to a group that was low. The point is that context matters. And even when findings generalize, they may not mean the same thing. We need to do a better job of contemplating —and explicitly addressing—contextual "twists" on generalizability.

Lest all these "demands" for doing more to recognize potential limits on the generalizability of findings seem daunting (even tempting some to "do something else": Yarkoni, 2022, p. 8), the fourth pointer to a way forward is to caution against hamstringing the field by demanding that all effects generalize broadly (Mitchell & Shivde, 2023). This would be a fool's errand, given that there is no reason to expect that any given sample on which an effect was observed is "normative" or informs a "default" human condition (Prather, 2023).

There is nothing wrong with pursuing effects that may hold for one group or under one set of conditions but not generalize beyond them. There *is* something wrong with not explicitly signaling these expectations, however. Absent tests for the generalizability of our effects, we must be loud and clear about reasons to expect when, where, and to whom they might generalize, and when, where, and to whom they might not. In short, as directed by Mitchell and Shivde, be honest.

By way of a final comment, I emphasize that there is no one way to "do" psychological science. Just as there is not a "normative" human (Prather, 2023), nor is there a "normative" way do to psychological science research. We may recognize the desirability of moving to a more model-centric approach (Devezer & Buzbas, 2023). Yet just as "practical relevance" (Otgaar et al., 2023) is not for everybody, everywhere, so too is it the case that a model-centric approach is not for everybody or for every phenomenon, or for every stage of a research endeavor. Rather than a one-size-fits-all approach, what we need is a very big tent—a tent that shelters all strong science, of all types. That said, psychological science is sufficiently mature that it seems reasonable and appropriate to charge "admission" to the tent—if you want to stand under it, you must be authentic. If we do not seriously reflect on the authenticity of the work we do, then as Mitchell and Shivde (2023) lament, we run the risk of going the way of alchemy.

Acknowledgments

Preparation of this article was supported by NIH grant HD094716, and by Emory College of Arts and Sciences.

References

Anvari F, & Lakens D (2021). Using anchor-based methods to determine the smallest effect size of interest. Journal of Experimental Social Psychology, 96, 104159. 10.1016/j.jesp.2021.104159

Bauer PJ (2023). Generalizations: The grails and the gremlins. Journal of Applied Research in Memory and Cognition.

Devezer B, & Buzbas EO (2023). Rigorous exploration in a model-centric science via epistemic iteration. Journal of Applied Research in Memory and Cognition.

Eich E (2014). Business not as usual. Psychological Science, 25(1), 3–6. 10.1177/0956797613512465 [PubMed: 24285431]

Fife D (2020). The eight steps of data analysis: A graphical framework to promote sound statistical analysis. Perspectives on Psychological Science, 15(4), 1054–1075. DOI: 10.1177/1745691620917333 [PubMed: 32502366]

Henri V, & Henri C (1895). On our earliest recollections of childhood. Psychological Review, 2, 215–216.

Henri V, & Henri C (1898). Earliest recollections. Popular Science Monthly, 53, 108–115.

Lindsay DS (2015). Replication in Psychological Science. Psychological Science, 26(12), 1827–1832. 10.1177/0956797615616374 [PubMed: 26553013]

Miles C (1893). A study of individual psychology. American Journal of Psychology, 6, 534–558.

Mitchell KJ, & Shivde G (2023). Generalizability in psychology research: Beware the Grinch. Journal of Applied Research in Memory and Cognition.

Müller-Lyer FC (1889). Optische Urteilstäuschungen. Archiv für Physiologie Suppl. 1889: 263–270.

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, … Yarkoni T. (2015). Promoting an open research culture. Science, 348(6242), 1422–1425. 10.1126/science.aab2374 [PubMed: 26113702]

Otgaar H, Riesthuis P, Neal TMS, Chin J, Boskovic I, & Rassin E (2023). If generalization is the grail, practical relevance is the Nirvana: Considerations from the contribution of psychological science of memory to law. Journal of Applied Research in Memory and Cognition.

Peterson C (2023). Gremlins in childhood amnesia research. Journal of Applied Research in Memory and Cognition.

Prather RW (2023). A new path: Why we need critical approaches to cognitive and psychological sciences. Journal of Applied Research in Memory and Cognition.

Simmons JP, Nelson LD, & Simonsohn U (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359–1366. DOI: 10.1177/0956797611417632 [PubMed: 22006061]

Sternberg RJ, & Gordeeva T. (1996). The anatomy of impact: What makes an article influential? Psychological Science, 7(2), 69–75. doi:10.1111/j.1467-9280.1996.tb00332.x

Yarkoni T (2022) The generalizability crisis. Behavioral and Brain Sciences, 45, e1: 1–78. doi:10.1017/S0140525X20001685