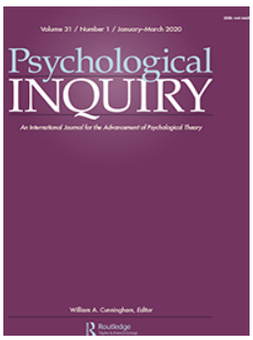

Routledge
Taylor & Francis Group

Psychological Inquiry

An International Journal for the Advancement of Psychological Theory

ISSN: 1047-840X (Print) 1532-7965 (Online) Journal homepage: https://www.tandfonline.com/loi/hpli20

Unjustified Generalization: An Overlooked Consequence of Ideological Bias

Yoel Inbar

To cite this article: Yoel Inbar (2020) Unjustified Generalization: An Overlooked Consequence of Ideological Bias, Psychological Inquiry, 31:1, 90-93, DOI: 10.1080/1047840X.2020.1724758

To link to this article: https://doi.org/10.1080/1047840X.2020.1724758

Published online: 09 Mar 2020.

PSYCHOLOGICAL INQUIRY
2020, VOL. 31, NO. 1, 90–93
https://doi.org/10.1080/1047840X.2020.1724758

Routledge
Taylor & Francis Group

Unjustified Generalization: An Overlooked Consequence of Ideological Bias

Yoel Inbar

Department of Psychology, University of Toronto, Toronto, Canada

Clark and Winegard (this issue) argue that researchers' (liberal) ideology is a threat to the validity of social-psychological research. They describe three ways in which researchers' ideological commitments can undermine the validity of social science: exaggerating ideologically-congruent small effects, ignoring ideologically-incongruent alternative hypotheses, and strategically framing findings in the most ideologically-appealing way. These threats to validity are important and should be taken seriously. But they seem minor in comparison to a more fundamental threat to validity that would remain even if we addressed them: unjustified generalization from laboratory studies (often, but not always, experiments) to the real-life situations that inspired them (and that they are ultimately meant to explain). In fact, this threat to validity is so common that it can be hard to notice.

Unjustified generalization is not limited to ideologically-relevant findings. But then, exaggerating small effects, overlooking alternative hypotheses, and selective framing are not limited to ideologically-relevant findings either. All of these practices are threats to validity and that distort the interpretation of findings in some way and often stem from researcher motivations. Those motivations might simply be self-serving (e.g. publishing in more prestigious outlets or capturing public attention), but they might also ideological or moral (e.g. supporting an ideological position that the researcher thinks is correct). Or, if researchers believe both that an ideological position is ought to be promoted and that work promoting it is more likely to be published, motivations may be a mix of morality and self-interest.

Unjustified generalization, then, is one of a number of threats to validity that interact with researcher ideology in a potentially troublesome way. In the case of unjustified generalization in particular, it seems very plausible that the more the explanation is ideologically appealing, the less motivated social psychologists are to think about the gap between the laboratory results and the real-world phenomena. One example of this may be the laboratory research demonstrating stereotype threat.

Stereotype Threat: A Case Study

Stereotype threat is the idea that members of a negatively-stereotyped minority group might underperform relative to their true ability on a task (such as a standardized test) because anxiety about confirming negative stereotypes impairs performance (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). Clark and Winegard cite this as an example of an exaggerated ideologically-congruent finding and point to problems with the stereotype threat literature such as publication bias and small effect sizes. But assume for the moment that stereotype threat using a specific experimental paradigm in the lab is 100% replicable with a large effect size. Does that tell us much about real-world disparities in test performance between social groups? By design, it cannot. It can only tell us that when using specific, carefully selected items, in specific, carefully designed experimental settings, a stereotype threat effect exists. Generalizing from the lab to the field requires an extra step of construct validation to show that the laboratory paradigm actually captures all the relevant psychology of the real-world situation it is intended to represent (Cronbach & Meehl, 1955). Without this step, any explanation of actual test disparities based on laboratory studies is just a hopeful speculation. Indeed, the more stereotype threat studies resemble the high-stakes testing they are meant to represent, the smaller stereotype threat effects appear to be (Shewach, Sackett, & Quint, 2019).

This is by no means specific to research on stereotype threat. Many social psychological experiments are existence proofs that under certain tightly controlled conditions, X can cause Y. Such experiments are simply not designed to answer the question of how much of a real-world phenomenon of interest (putatively represented by Y) is caused by X. This is not a problem as long as the existence proof, or the illumination of a psychological mechanism *per se*, is what is interesting (Mook, 1983). Sometimes that is true, but often it is not. Often the "hook" for the research is some real-world issue or problem, and the experiment is implicitly or explicitly meant to speak to it. This is certainly the case with stereotype threat research. If it were a phenomenon that existed only in specific, circumscribed laboratory settings, it would be of much less interest. A great appeal of the stereotype threat research (especially to liberals) is that it purports to speak to real-world achievement gaps between members of majority and minority groups. In an amicus curiae brief to the US Supreme Court, for example, leading stereotype threat researchers argued that "standardized test scores and grades often underestimate the

CONTACT Yoel Inbar yoel.inbar@utoronto.ca 1265 Military Trail, Toronto, ON, M1C 1A4.

true academic capacity of members of certain minority groups" (Aronson et al., 2015, p. 4).

On their own, however, laboratory demonstrations of stereotype threat cannot actually explain real-world achievement gaps; even statistically robust effects only disconfirm the much more limited null that *under no circumstances* will stereotype awareness affect the performance of minority group members (see Yarkoni, 2019). Perhaps that null is worth disconfirming. But the stereotype threat hypothesis (broadly construed) is on its face plausible enough that one might wonder whether that is even necessary. Perhaps one would do just as well to defend the stereotype threat hypothesis logically instead of empirically, and go straight to answering the question people actually care about: whether stereotype threat causes significant performance decrements *in situations that actually matter.*

This real-world-first approach mitigates one of the worst consequences of experimentation-first: premature certainty. Because experiments allow researchers the freedom to choose all aspects of the participants' experience in the study, there is a high likelihood that (at least for somewhat plausible hypotheses) a skilled researcher will calibrate the experiment precisely so as to produce a statistically significant effect that is also severely limited in generality. As McGuire (1983, p. 15–16) put it, "It can be taken for granted that some set of circumstances can be found to confirm any expressible relationship, provided that the researcher has sufficient stubbornness, stage management skills, resources, and stamina sooner or later to find or construct a situational context in which the predicted relationship reliably emerges." Having successfully demonstrated the phenomenon in (perhaps many) laboratory experiments, researchers then feel an unfounded certainty that this phenomenon must obtain in the real-world situation of interest, as well as the lab.

It is worth considering what a program of research on stereotype threat would have looked like had it not focused first and foremost on demonstrating the phenomenon in the laboratory. Progress would certainly have been slower. It is obviously more difficult to intervene in high-stakes testing than to run laboratory experiments with 40 undergraduates. But if the question of interest is whether performance on high-stakes tests is affected by stereotype threat, ought one not start by carefully investigating performance on high-stakes tests, rather than by constructing a laboratory paradigm that may be unrepresentative of the actual situation of interest in many ways? Such an approach might start first by careful measurement of real-world phenomena, and by making predictions about observed relationships based on the theory. For example, to the extent that minority group members experience the preconditions of stereotype threat, their future performance ought to be underpredicted by test scores. (This approach, known as "differential prediction," is a standard way of assessing possible test bias; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014.) Had this been the first step, researchers might have encountered negative evidence—i.e. a lack of real-world evidence of differential prediction consistent with stereotype threat—at the outset (Cullen, Hardison, & Sackett, 2004). Or perhaps researchers might have tested a real-world intervention thought to reduce stereotype threat—e.g. asking test-takers for demographic information after the test rather than before it. Again, had they done so, they might have found earlier on in the research process that the intervention failed to reduce the test-score gap as predicted (Stricker & Ward, 2004).

Observational and field research has its own problems of interpretability, in particular for causal questions. However, it also is an important sanity check. In doing such research one might run up against inconvenient failures of theoretical prediction that are harder to "fix" by altering the experimental design. These early inconsistent observations are likely to encourage researchers to rethink their predictions. This approach of going to the field and allowing interventions to fail has proven to be extremely productive in other disciplines such as development economics (Jayachandran, 2019).

If instead, researchers first run hundreds of lab experiments confirming stereotype threat (Aronson & Dee, 2012), by the time they venture out of the lab, they will have convinced themselves that stereotype threat must have significant effects in real-world settings (how could hundreds of lab experiments be wrong?) and they will thus be very reluctant to take no for an answer. This reluctance to take no for an answer means that when a large field experiment in actual high-stakes testing fails to confirm the predictions of stereotype threat (e.g. Stricker & Ward, 2004) the main finding is discounted and selective re-analyses are done to generate a theory-consistent finding (e.g. Danaher & Crandall, 2008; see Sackett & Ryan, 2012).

Alternatively, operationalizations can mutate such that the relationship to the original laboratory research seems more like an analogy than an application. Support for the predictions of stereotype threat theory is claimed from the effects of multifaceted interventions such as values affirmations (Cohen, Garcia, Apfel, & Master, 2006), social belongingness (Walton & Cohen, 2007), or theories of intelligence (Aronson, Fried, & Good, 2002); and outcome measures very different from the original lab studies (e.g. course grades or cumulative undergraduate GPA). These interventions and outcomes are argued to tap the same underlying psychology, but this seems based more on faith than evidence. At the extreme, the logic becomes almost entirely circular: if the intervention boosted the performance of minority group members, it must have affected stereotype threat, and the prevalence of stereotype threat is demonstrated by the fact that the intervention studies "worked." It is doubtful that standards of evidence would be so relaxed without the certainty-bolstering effects of hundreds of confirmatory lab studies.

It's Not Just Stereotype Threat

I have focused on stereotype threat in particular, but the problem of unjustified generalization is old and widespread, going back to the classics of social psychology

(Brannigan, 2004). For example, Milgram's (1974) obedience experiments, which were inspired by Nazi war crimes, say little-to-nothing about Nazi war criminals, many of whom appear to have participated not only willingly but enthusiastically (Goldhagen, 1997).

Beyond excessive confidence in simplistic causal explanations for complicated real-world phenomena, unjustified generalization can also have more subtle consequences. In focusing researchers on laboratory effects of unknown real-world relevance, it can lead to arguments about numbers that are simply irrelevant to the actual question of interest. For example, Clark and Winegard mention that a recent meta-analysis shows that the relationship between implicit attitudes and behavior is $r = 0.09$ (Forscher et al., 2019). But this is just the average effect size for the measures social psychologists happened to use for those studies, which might have been chosen for any number of reasons: maximizing effect size, rhetorical impact, precedent in the literature, convenience, etc. Here is how Forscher et al. described the measures included in the meta-analysis (p. 529, citations have been omitted): "behavioral tasks involved a wide range of outcomes, such as seating distance from a Black or White confederate, willingness to participate in a hypothetical beer pong game, intentions to drink in the future, reported chocolate consumption, and intentions to vote for gay and lesbian civil rights referenda." The average of the correlation between implicit attitudes and Black-White seating distance, self-reported chocolate consumption, and hypothetical willingness to play beer pong (among other things) has no obvious real-world meaning (see also Simonsohn, 2015), and so is not at all relevant to the question of the real-world effects of implicit bias. Neither, however, is the average effect size for 200 studies examining the relationship between implicit attitudes and seating distance from a Black confederate—unless the work of establishing a link between seating distance and real-world discrimination has been done.

Conclusion

I am not arguing that laboratory experiments have no value. Existence proofs can be useful, and often we want to understand psychological process rather than explain a particular real-world phenomenon. When this is the case, laboratory experiments are essential. But when it is not, we should be much more careful about generalizing from the lab to the field. Even interventions that seem strongly theoretically justified based on laboratory research (and intuitively plausible) often fail in practice (for example, see Glewwe, Kremer, & Moulin, 2009; Goswami & Urminsky, 2019; Jung, Perfecto, & Nelson, 2016). In principle, social psychologists acknowledge that generalizing beyond the lab requires extensive research. In practice, they are often perfectly happy to point to a (deliberately) simplified, reductive lab finding to explain a complex social phenomenon. This problem is likely to be particularly severe for ideology-friendly findings. This is an ironic state of affairs for a discipline whose central tenet is that behavior differs dramatically across situations. We would do well to take this tenet seriously and start with more real-world observation and less (confirmatory) experimentation, particularly when we are trying to explain important real-world phenomena.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Joint Committee on Standards for Educational and Psychological Testing.

Aronson, J., & Dee, T. (2012). Stereotype threat in the real world. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (p. 264–278). Oxford, UK: Oxford University Press.

Aronson, J., Dweck, C. S., Erman, S., Good, C., Inzlicht, M., Logel, C., … Yeager, D. (2015). *Brief of experimental psychologists as amici curiae in support of respondents* (pp. 579). US: Fisher v. University of Texas at Austin.

Aronson, J., Fried, C., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*(2), 113–125.

Brannigan, A. (2004). *The rise and fall of social psychology: The use and misuse of the experimental method*. New York: Aldine de Gruyter.

Cohen, G., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*(5791), 1307–1310. doi:10.1126/science.1128317

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi:10.1037/h0040957

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*, *89*(2), 220–230. doi:10.1037/0021-9010.89.2.220.

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings reexamined: A reply. *Journal of Applied Social Psychology*, *34*, 1656–1663. doi:10.1111/j.1559-1816.2008.00362.x

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522–559. doi:10.1037/pspa0000160

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics*, *1*, 112–135. doi:10.1257/app.1.1.112

Goldhagen, D. J. (1997). *Hitler's willing executioners: Ordinary Germans and the Holocaust*. New York, NY: Vintage.

Goswami, I., & Urminsky, U. (2019). *No substitute for the real thing: The importance of in-context field experiments in fundraising*. Retrieved from http://home.uchicago.edu/~ourminsky/Goswami_Urminsky_No_Substitute.pdf.

Jayachandran, S. (2019). When a disappointment helped lead to a Nobel Prize. *The New York Times*. Retrieved from https://www.nytimes.com/2019/11/29/business/economics-nobel.html.

Jung, M. H., Perfecto, H., & Nelson, L. D. (2016). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. *Journal of Marketing Research*, *53*(3), 354–368. doi:10.1509/jmr.14.0238

McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovation and reform in psychological research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 16, pp. 1–47). New York: Academic Press.

Milgram, S. (1974). *Obedience to authority*. New York: Harper and Row.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379–388. doi:10.1037/0003-066X.38.4.379

Sackett, P. R., & Ryan, A. M. (2012). Concerns about generalizing stereotype threat research findings to operational high-stakes testing. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (p. 249–263). Oxford, UK: Oxford University Press.

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514–1534. doi:10.1037/apl0000420

Simonsohn, U. (2015). "*The*" *effect size does not exist.* Retrieved from http://datacolada.org/33.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. doi:10.1006/jesp.1998.1373.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. doi:10.1037/0022-3514.69.5.797.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*(4), 665–693. doi:10.1111/j.1559-1816.2004.tb02564.x

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*(1), 82–96. doi:10.1037/0022-3514.92.1.82

Yarkoni, T. (2019). *The generalizability crisis.* Retrieved from 10.31234/osf.io/jqw35.