Replication or Generalizability?
How Flexible Inferences Uphold Unfounded Universal Claims

Moin Syed
Department of Psychology, University of Minnesota

Paper presented at the Metascience 2023 Conference, Washington D.C.
10 May 2023

Version Date: May 8, 2023

Author Note

The impetus for exploring these issues came about following an email exchange with my colleague Matt McGue, which goes to show that email is not 100% evil. This is a rough extended script of a presentation, and thus does not follow all conventional citations patterns or formalities of scientific writing. It has not been peer reviewed. It could turn into a proper paper someday, however, so I welcome any comments, moin@umn.edu. Slides and this paper available at https://osf.io/6xsmc/

Abstract

Metascientists commonly distinguish between tests of replicability, reproducibility, robustness, and generalizability. Whereas these distinctions are sensible, they are not often the subject of in-depth definitional exploration. This paper examines the relation between two of these—replication and generalizability—seeking to clarify their differences in the context of sample diversity. Following a brief discussion on the definitions of these terms, I advance four assertions: 1) discussions of replication assume researchers adhere to the Mertonian norm of disinterestedness, 2) researchers are motivated to maintain claims associated with their ontological stance, most commonly universalism, 3) despite universal claims, failures of replication are perceived to have greater epistemic consequences than lack of generalizability, and 4) researchers are motivated to reframe threats to replicability as limits of generalizability. I then use behavioral genetics research on polygenic score prediction of educational attainment as a case study to illustrate the assertions. Finally, I close with some brief recommendations for how to move forward.

Keywords: replication, generalizability, diversity, reproducibility, behavior genetics, polygenic score

As metascience has grown in popularity, so too has the need for clear, consistent, metascientific language. Although terms such as replicability and generalizability have long been part of scientific nomenclature, their meanings have not historically been subject to scrutiny by practicing researchers. Rather, they tend to be understood as intuitive folk concepts—foundational ideas that everybody generally understands but rarely thinks about deeply.

In psychology and related fields, the replication crisis and associated open science movement has resulted in widespread efforts to not only develop a consistent terminology, but also a coherent understanding of the issues and practices that underlie our terms. A broadly disseminated framework developed by The Turing Way (2021) makes clear distinctions among replicability, reproducibility, robustness, and generalizability by setting up a 2x2 matrix with one dimension pertaining to whether the same data are being used and the other pertaining to whether the same analysis is being used (Figure 1).

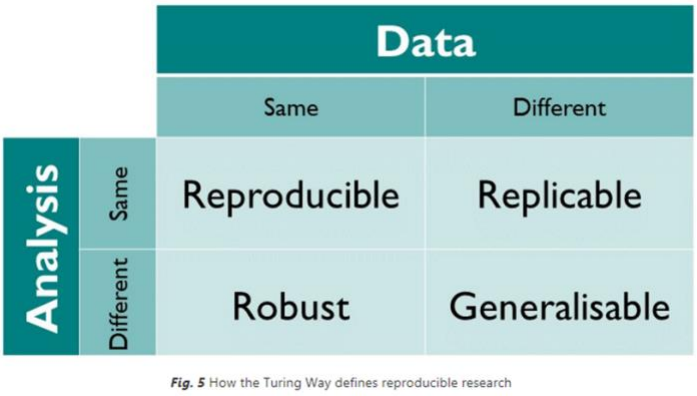

Figure 1. Framework for understanding reproducible research, from The Turing Way Community (2021)

Replicability and reproducibility are particularly prone to confusion and are often used interchangeably, a situation exacerbated by the Open Science Collaboration's (2015) landmark replication project being titled, "The reproducibility of psychology science," an unfortunate situation that was rectified in the subsequent, "Investigating the replicability of preclinical cancer biology" (Errington et al., 2021).

The purpose of this paper, however, is to focus attention on a different comparison in The Turing Way framework, namely, how to think about the relation between replications and generalizations. At the broadest level, I seek to a) use sample diversity to demonstrate that this distinction is not so clear-cut and b) argue that the distinction is further muddied by researchers' motivated behavior. First, I provide a very brief summary of how replication and generalizability tend to be defined. Second, I advance a series of assertions about the relation between replication and generalizability[1] that are situated within the psychology and sociology of science. Third, I use the behavioral genetic research program on polygenic scores for educational attainment to illustrate my assertions. Fourth,

[1] Grammarians will notice that I do not keep parallel forms of replication and generalization through the paper. Whereas both replication and replicability are typically used in the literature, generalizability is the most common form (vs. generalization), and thus that is what I tend to use regardless of whether the comparative referent is replication or replicability.

and finally, I provide some recommendations on how to refine our collective thinking to improve the inferential process of our research.

On Definitions of Replication and Generalizability, Briefly

Unsurprisingly, the simplicity of The Turing Way 2x2 framework betrays the complexity of each of the cell entries, particularly for replication and generalizability. Psychologists have long distinguished between two types of replication: *direct (or exact) replication*[2], which seeks to repeat the procedure of the target study as close as possible, and *conceptual replication*, which seeks to test the same target hypothesis via some number of changes in the study procedure. Psychologists have long valued conceptual replications over direct replication, arguing that the former provide greater insights into the underlying phenomena and that the latter are essentially mundane and unnecessary once a claim has been established (Crandall & Sherman, 2016), often via a single study. In the wake of the replication crisis, the value of direct replications increased, with proponents arguing that it is necessary to first develop confidence in the reliability of a finding before seeking to broaden it out via conceptual replication (Zwann et al., 2018).

It should be clear that direct and conceptual replications are quite different in both substance and inferential currency. Conceptual replications are not simply the same analysis using different data, as implied by The Turing Way 2x2, but rather represent different analyses with different data, because the study conditions will often be substantially different from the targets. In this way, conceptual replication occupies the same cell as generalizability. Indeed, any discussion on the value of conceptual replication will invoke some aspect of generalizability as the motivation of the study (Lykken, 1968; Zwann et al., 2018). Additionally, both the direct replication → conceptual replication and replication → generalizability sequences are described as the optimal approach to move a research program along (Melhuish & Thanheiser, 2018; Sikorski & Adreoletti, 2023).

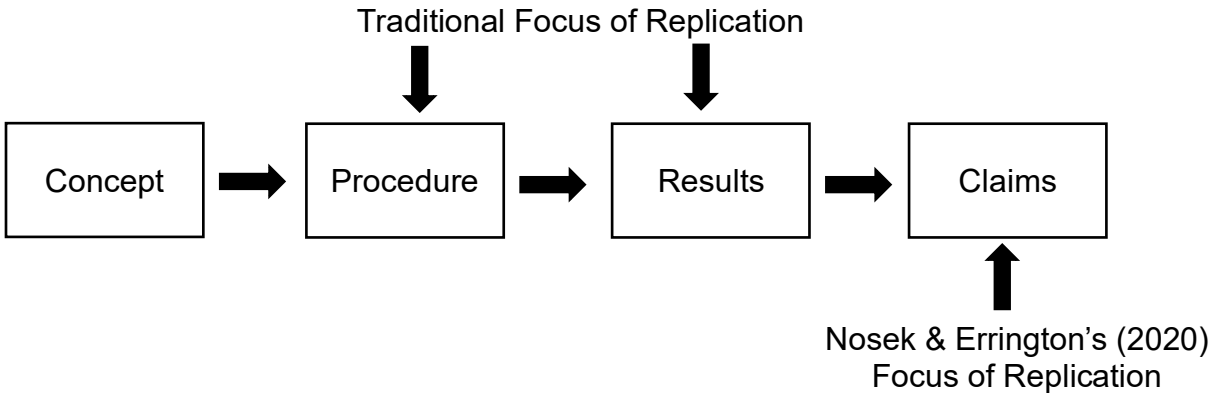

Figure 2. Shifting the focus on replications from the procedures and results to the claims made.

[2] Pedantic and otherwise annoying people are often quick to point out that no replication in psychology can ever be "exact," because at minimum the historical conditions will be different in the replication. As you will see if you keep reading, this argument is quite ironic. Such objections are one reason to favor the use of "direct" replication and be clear that the intent is to follow the original as closely as possible. Lykken (1968) seemed to anticipate this objection, making a distinction between *literal* replication and *operational* replication, the latter of which is most closely associated with current use of direct replication: "…duplicate exactly just the sampling and experimental procedures given in the first author's report of his research." (p. 155). (Lykken also discussed *constructive* replication, akin to today's *conceptual* replication.)

This terminological messiness seems to be part of the motivation for why Nosek and Errington (2020) articulated a redefinition of replication: "According to common understanding, replication is repeating a study's procedure and observing whether the prior finding recurs[7]. This definition of replication is intuitive, easy to apply, and incorrect." (p. 2). The crux of their argument is that, rather than defining replication in relation to a study's *procedures*, we should define replication in relation to a study's *claims* (Figure 2). Replication, then, "…is a study for which any outcome would be considered diagnostic evidence about a claim from prior research." (p. 2).

As part of this redefinition, Nosek and Errington (2020) do away with the distinction between direct and conceptual replication, because the distinction over-emphasizes the procedures vs. the claims. If a replication is any test that is considered diagnostic of the truth of a particular claim, then generalizability is any test that seeks to test the boundary conditions of the claim. As I noted previously, in common usage the distinction between conceptual replications and tests of generalizability are not at all clear. Nosek and Errington's (2020) definitional work helps resolve this tension. In addition to clearing up terminological confusion, this definition of replication has the benefit of drawing greater attention to the claims that researchers make, as well as the connections between their claims and their evidence. This is a crucial point that I will return to in a bit.

Nosek and Errington (2020) also discussed the dynamic between replication and generalizability, highlighting how it changes as a research program matures (Figure 3). Early on, when little is known, claims will have insufficient precision and so the potential generalizability space is large relative to the replication space. Over time, the relation reverses, where the replication space is much larger relative to generalizability space.

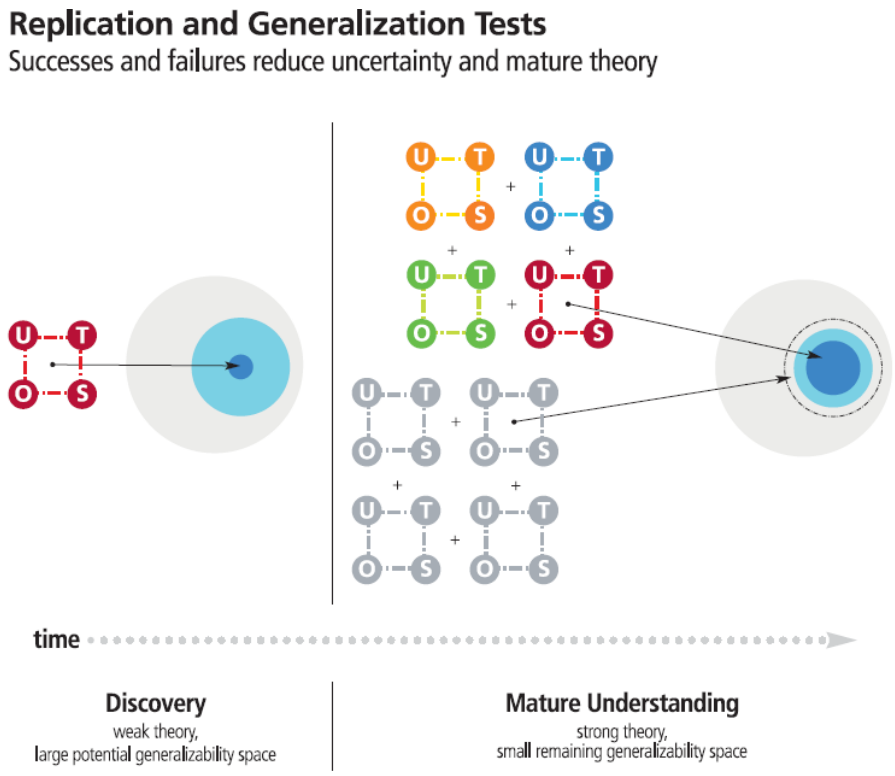

Figure 3. From Nosek and Errington (2020), showing how the relation between replication and generalizability changes as a research program matures.

Interestingly, this particular point about the informativeness of replication over time may not be dependent on defining replication with respect to claims vs. procedures. In positioning their view as against the mainstream metascience literature, Devezer and Buzbas (2022) maintain a procedural definition of replication: "It is called an exact replication, if this experiment shares the same assumed model, method, and data structure as the original but differs in background knowledge and data values." (p. 2). A major point of departure is that they explicitly reject the widespread notion that replication is a cornerstone of scientific progress (see also Sikorski & Adreoletti, 2023), arguing that both true discoveries can yield failures to replicate, and that highly replicable findings can be false. The first argument converges with Nosek and Errington (2020), that the utility of replications changes as a research program matures and precision increases. In other words, not all studies are ready to be replicated (see also Scheel, 2022). The second argument, that highly replicable findings can be false, is obviously true, and possibly widespread. For example, both text-based (Youyou et al. 2023) and empirical (Soto, 2019) estimates suggest that replication rates are very high in personality psychology, perhaps the highest in all of psychology. Personality psychology relies heavily on observational data, and thus there are clearly many confounds and a substantial degree of "crud" (Meehl, 1984; Lykken, 1968) that can make spurious results appear replicable and true (Soto, 2021). Replication is neither necessary nor sufficient for diagnosing the truth of claims, and thus perhaps it should not have the high privilege status that it does.

Indeed, psychology has long prioritized internal validity, of which replication is a part, over external validity, of which generalizability is a part. Esterling et al. (2023) argued for the need to elevate external validity and see it as a core component of *generalized causal claims,* or quite simply a generalized claim about cause and effect (see Deffner et al., 2021 for a similar argument). These are precisely the kinds of claims that psychologists—and those from other social science disciplines—tend to make. Esterling et al. also do some redefinition work, arguing that the traditional definition of external validity, that a claim is valid if it generalizes across settings, does not suffice. Rather, they argue that external validity should be evaluated with respect to whether specifications have been made as to *how* and *why* an effect generalizes; that these features are part of the causal claim.

In proclaiming a "generalizability crisis," Yarkoni (2021) identified myriad study features that social scientists rarely examine for the generalizability of their claims. These include, but are not limited to, measurements, research site, or sample characteristics. This highlights how generalizability is even more diffuse a concept than replication, and thus there is a need to be clear about the dimensions of generalizability that are the focus of discussion. In what remains, my focus is on sample diversity broadly, and particularly on diversity with respect to race, ethnicity, culture, and nation. The lack of sample diversity has long been discussed in psychology, yet we have seen little progress. As I will argue, there are reasons for why this is the case.

Four Assertions on Replication and Generalizability

With the definitional issues in mind, I advance the following assertions, which are the central purpose of this paper:

Assertion 1: Discussions of replication assume researchers adhere to the Mertonian norm of disinterestedness. In reading the literature on replication, much of it seems to assume that researchers are disinterested truth-seekers who act for the good of science. Although this has long been questioned, researchers' motivated reasoning became plain from the beginning of the replication crisis. Some unknown number of researchers are motivated to maintain their belief in

their original results even as the evidence to the contrary mounts (e.g., Baumeister, 2019; Strack 2016). Another way this is evident is in the case where a correction is made to an article, yet the authors state that it "does not impact the substantive conclusions," as though claims are not, in fact, reliant on the underlying evidence. Any discussion about replication and its role in science must build in, not just acknowledge, the fact that researchers do not "play fair" according to some neutral rules (see also Lebel & Peters, 2011). That said, not all forms of self-interest are this extreme; some are more subtle commitments to particular views.

Assertion 2: Researchers are motivated to maintain claims associated with their ontological stance, most commonly universalism. Mainstream psychological researchers adopt a universalist perspective, wherein their samples are intended to represent all of humanity. The field relies heavily on convenience samples of relatively privileged individuals (White, middle-to-high SES, from USA/Western Europe) yet makes claims about general processes universal to humans—generalized causal claims, as described by Esterling et al. (2023). That researchers are motivated to maintain this stance is precisely why there has been a long procession of commentaries bemoaning the lack of diversity in samples used in the field for the last 50 years, yet we have seen little change (Arnett, 2008; Graham, 1992; Guthrie, 1976; Hartmann et al., 2013; Henrich et al., 2010; Moriguchi, 2022; Roberts et al., 2020). It is fairly obvious that mainstream psychological researchers want to make universal claims without marshalling the sort of evidence that would be needed to support them.

It is important to note that motivation to maintain universalism is only evident within *mainstream* research. What, exactly, constitutes "mainstream research" is tough to say, but there are at least two examples that highlight what is not included in this category. Multiple lines of evidence illustrate that articles that rely on samples from outside the U.S. and Western Europe are more likely to include the country name in the title (Cheon et al., 2020; Kahalon et al., 2022) and that within the U.S., articles focused on racial/ethnic minorities are more likely to include the participants race/ethnicity in the title compared with studies with White samples (Cheon et al., 2020; Roberts & Mortenson, 2022). Whether this behavior is self-chosen or enforced by journal editors is a complex discussion for another day, but it highlights how some research (e.g., with White participants from the U.S. and Western Europe) is construed as universal whereas other research is not.

Assertion 3: Despite universal claims, failures of replication are perceived to have greater epistemic consequences than lack of generalizability. Given the claims of universalism, both failures to replicate past findings and evidence of limits of generalizability should be seen as epistemic threats. Failures of replication are clearly seen this way, because of their ostensible diagnostic value as to whether or not a finding is "true" (but again, see Devezer & Buzbas, 2021). This is the reason why a sudden rash of failures to replicate past findings were collated under the heading of "replication crisis," and the reason for all of the resultant actions to improve the situation (including this conference). Limits of generalizability do not have the same weight, typically relegated to half-hearted apologies in the Limitations sections of articles (Clarke et al. 2023). Despite the efforts of some—cultural and ethnic minority psychologists with respect to sample diversity (e.g., Hartmann et al., 2013) and people like Yarkoni (2021) about study design features—there has not been a major reckoning in the field about generalizability. As Tiokhin et al. (2019) put clearly, what we tend to see is, "brief caveats that acknowledge the unrepresentativeness of participants, cite the WEIRD paper, and go about business as usual" (p. 2). Thus, not only do mainstream psychological researchers want to make universal claims without marshalling the sort of evidence that would be needed to support them, they don't care that they are failing in this regard when it is pointed out to them. In short, we take replication seriously, but are happy to ignore generalizability.

Assertion 4: Researchers are motivated to reframe threats to replicability as limits of generalizability. Given a) the perceived epistemic threat of replication, b) the general shrug about lack of generalizability, and c) that there are not always clear distinctions between what counts as a test of replication vs. test of generalizability, researchers are motivated to reframe threats to replicability as limits of generalizability. This practice is clearly evident in two of the prominent talking points against needing to take the failures of replication seriously: hidden moderators (Stroebe & Strack, 2014) and contextual sensitivity (Van Bavel et al., 2016)). Once again, these claims need to be understood in the context of the general tendency to make universal claims in the original reports. Thus, these defenses are made *post-hoc*, after the failures to replicate are known. Such post-hoc defenses cannot be considered credible (Esterling et al. (2023), and neither of these specific defenses has fared well when examined empirically (Ebersole et al., 2020; Inbar, 2016). This assertion is really just a more specific version of *Assertion 1*, that researchers are motivated parties and do not adhere to what should be normative standards of evaluating evidence.

Taken together, the four assertions suggest that researchers can play fast and loose with labeling a test as related to replicability or generalizability. The use of the term *conceptual replication* and the heavy focus on study procedures vs. study claims enables both the confusion about the terms (and underlying concepts) and the opportunity for researchers to reframe tests as needed to protect their claims. Whether or not you fully agree that replication should be defined in relation to claims vs. procedures, anyone who is serious about evidence would agree that we should heavily scrutinize research claims and the evidence that is provided in support of them.

Most of us would probably endorse the idea that our scientific pursuits should result in *cumulative knowledge* (Hedges, 1987). This is not to say that we engage in a purely linear process of accumulating facts—our work involves starts, stops, and regressions—but that we should ultimately emerge feeling that we have increased our understanding of the target phenomenon. This is generally true for basic knowledge about how the world works, but is especially true if we seek to use that knowledge for application to people's lives. To illustrate my four assertions, and why I think they matter, I provide a brief case study examining the behavioral genetic research on polygenic scores for educational attainment.

Case Study: Polygenic Scores in Behavioral Genetics Research on Educational Attainment

Behavioral genetics is often highlighted in discussions about the replication crisis for two reasons: 1) they were at the forefront of realizing they had a major replication problem with so-called "candidate gene" studies, which sought to link a single genetic polymorphism to some outcome (Hewitt, 2012), and 2) they quickly mobilized to rectify the problem by creating consortia to collaborate on large-sample genome-wide association studies (GWAS), which link variation across the entire genome with a phenotype (Abdellaoui et al., 2023). This methodological move occurred alongside a major conceptual one, namely the recognition that there are many genes involved in phenotypic expression, each exhibiting a small effect, rather than singular genes that could be reliably linked to outcomes (Chabris et al., 2015). An extension of this approach is the use of polygenic scores[3], which are weighted sums of trait-associated alleles for each individual that can be used to make individual predictions (Kullo et al., 2022).

[3] Polygenic scores are also inconsistently labeled polygenic risk scores and polygenics indexes, and are represented by the acronyms PGS, PRS, and PGI. To the best of my knowledge, these are all different terms for the same score (jangle!)

Because of the need for very large samples (hundreds of thousands, if not millions of participants) it was necessary to pool data across multiple sources. However, few phenotypes are consistently measured across samples, and thus there are limited opportunities to pool data. Behavioral genetics researchers identified a socially relevant phenotype that is almost always included in studies: educational attainment. Starting with Rietveld et al. (2013), there have now been four successive GWAS of educational attainment, each increasing in sample size and analytic sophistication over the previous.

The line of research on polygenic score prediction of educational attainment serves as an ideal example because a) approximately the same research design has been used across four studies, b) the same general research team has conducted the four studies, and c) the progression of the four studies has paralleled increasing recognition and discussion of the limitations of lack of sample diversity in GWAS. Although these particular features are unique and make the set well-suited to illustrate my core arguments, I maintain that there is nothing particularly unique to GWAS in the observations that I outline below. Generally, the same is likely to be seen in any progression of research in any social sciences that is not overly concerned with the implications of sample diversity.

The first GWAS of educational attainment was titled, "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment" (Rietveld et al., 2013). The paper reported that the sample was restricted to "Caucasian[4]" participants, and no analyses were conducted to examine the comparative predictive power of the resulting polygenic score for a different ancestry group. The analyses were conducted only with individuals of European ancestry, but the title indicates a general claim, and neither the abstract nor the discussion section constrains any of the claims with respect to sample diversity. This is rather typical of research in psychology and illustrative of my Assertion 2, that researchers seek to make universal claims that are not supported by the underlying data.

The second GWAS, "Genome-wide association study identifies 74 loci associated with educational attainment" (Okbay et al. 2016), looked very much like the first. Although the authors had ditched the "Caucasian" label for "European descent," there once again was no cross-validation in a different ancestry group, and title, abstract, and discussion all presented unconstrained claims. As the research had progressed since the first study, the claims of potential implications grew, "Our findings demonstrate that, even for a behavioural phenotype that is mostly environmentally determined, a well-powered GWAS identifies replicable associated genetic variants that suggest biologically relevant pathways" (p. 539).

With the third GWAS, "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals" (Lee et al., 2018), things begin to look a bit different. The discovery sample was once again restricted to individuals of European ancestry, but they examined the predictive power of the polygenic score in a new sample of African Americans. They found an R^2 attenuation of 85%, commenting that "this amount of attenuation is typical of what has been reported in previous studies" (p. 1115). Nevertheless, the title and abstract present unconstrained, universal claims, although the end of the discussion section cites the limited predictive power in other groups and thus, "the [polygenic] score will be most useful in research that is focused on samples of individuals with an European ancestry" (p. 1116). Notably, this was after a

[4] The label "Caucasian" has direct ties to scientific racism and is thus best avoided (Teo, 2009).

relatively extensive discussion on the implications of the results with respect to identifying genetic mechanisms associated with educational attainment and cognition, increasing power in randomized controlled trials, and furthering the study of gene-environment interactions.

The Lee et al. (2018) study represents a turning point in this line of research, both by empirically examining how well the findings from the European ancestry group hold up in a new group, and also by explicitly acknowledging that the findings likely have limited—rather than universal—applicability. This change is not terribly surprising, because as noted, alongside this line of research was the increasing recognition of the importance of sample diversity for GWAS (e.g., Martin et al., 2017). This change in the third study in the series is what makes this case study particularly useful, because there is a fourth study in the series so we can see exactly how the lessons learned from Lee et al. with respect to the role of sample diversity were evident in the subsequent study. Spoiler alert: they mostly were not.

Surprisingly, given what was learned in the previous Lee at el. (2018) study, in "Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals" (Okbay et al., 2022), there were no constraints indicated in the title or the abstract. Rather, the potential applicability was expanded in the abstract, wherein the authors claimed, "…polygenic index, explains 12–16% variance and contributes to risk prediction for ten diseases." (p. 437). Once again, the discovery sample was composed entirely of individuals of European ancestry. Once again, they examined the predictive power of the polygenic score in a new sample of African Americans. Once again, they found massive attenuation (85-89%) in the predictive power of the polygenic score[5]. However, the discussion section not only contains unconstrained claims, but there was also no mention at all of the limitation of the lack of predictive power in the African ancestry group. Rather, once again, there were broad claims about the applicability of the findings. One reason for this could be that this limitation is so well known as to be obvious, and thus does not need to be stated (Clarke et al., 2023)[6].

The lack of portability of polygenic scores across ancestry groups—in general, not just for educational attainment—is a problem that is widely known by anyone working in behavioral genetics. As demonstrated in the case study, there is increasing empirical attention to quantifying the lack of portability. These tests, which typically consist of polygenic scores derived from a European ancestry sample being used to predict outcomes in a different ancestry group (but see Saunders et al., 2022), are generally considered to be tests of generalizability. Do the findings from one group generalize to another? This is reasonable under the normative definitions of replication as repeating a study procedure and generalizability as expanding to new conditions.

If we think about replication as any test that is diagnostic of the original claim, however, things look quite different. The claims about polygenic score prediction of educational attainment are clearly universal claims. Across the four papers, there were no constraints on the universal claims made in the title, abstract, or body of the papers. The claims are not limited to individuals of European ancestry, but rather are unconstrained claims about all of humanity. This means that any similar test

[5] Note that in the main text of the paper they reported a different portability analysis than what was reported in Lee et al. (2018), but the supplemental materials included the comparable analysis, which is what I report here for consistency.
[6] It is also worth noting that this paper was published in *Nature Genetics*, which has a highly restrictive word count for the published paper, with most of the details in the supplement. Regardless, how authors choose to use the limited word count is revealing of where they assign importance.

conducted with a new sample—whatever their characteristics—is a test of replication, not a test of generalizability. The lack of predictive power in non-European ancestry groups are failures to replicate the original claim.

Most readers, especially those who work in behavioral genetics, are not likely to agree with that analysis. This is exactly what I would expect based on my four assertions. Researchers are motivated to maintain their universal claims (Assertions 1 and 2), view lack of replication as a more serious threat than lack of generalizability (Assertion 3), and thus reframe failures of replication as limits to generalizability (Assertion 4). Of course, this mapping only holds if one accepts a definition of replication as a diagnostic test of a claim. Putting aside the labeling of replication/generalizability, though, there is still a clear mismatch between the claims and the data, which at the end of the day is the more pressing concern.

Conclusions and Recommendations

I noted at the outset of the previous section that there was nothing particularly unique about the observations across the four educational attainment studies. The relation among those studies is somewhat unique, and allows for a direct examination of research progression with few confounds, but the observed pattern really isn't. Generalized causal claims are frequently made in psychology research yet the associated data rarely support them. Awareness and discussion of the lack of sample diversity in GWAS have seemed to increase (Wojcik et al., 2019), but actual diversification of samples has not (Fatumo et al., 2022). So too is the case for just about any area of psychology.

So what to do to improve the situation? Below I describe two suggestions, but you will note that one of them is *not* to be certain to include lack of sample diversity as a limitation. Doing so is not bad, per se, but limitations stated in articles can be understood as issues worthy of consideration, but not so serious as to revise the main claims or to be a death blow to the study itself (Clarke et al., 2023). The following suggestions take lack of sample diversity more seriously.

1. ***Make calibrated claims.*** Aligning claims with evidence is a simple change that does not require any additional resources to enact. Aligning, or calibrating, claims with evidence is something we should all seek to do (Vazire, 2018). This suggestion is similar to the proposal to add a "constraints on generality" section to discussion section of articles (Simons et al., 2017), but is meant to go a step further to think about constraints and calibrated claims from top to bottom (e.g., in the title and abstract). One of the major challenges is knowing what dimensions on which to calibrate. Especially early in a research program, such constraints may not be obvious until further study (Nosek & Errington, 2020). For example, the first two studies on educational attainment did not know for sure that the predictive power would attenuate in a different ancestry group (although one would have guessed, based on other evidence). The second two, however, did know that the results were ancestry specific, and thus the claims could have been adjusted accordingly. The reason that they did not is likely related to the most substantial challenge that this change requires: divesting from strong beliefs in universalism (Assertion 2). Having more constrained, calibrated claims is much more useful than holding strong with unfounded universal claims (Haeffel et al. 2023).

2. ***Build heterogeneity into our theories and methods***. This is obviously much more difficult. Note that this recommendation is not simply "diversify your samples." Whereas that is also probably a generally good idea, new inferential problems can arise when doing so (see Syed,

2021). Rather, this recommendation is more aligned with Esterling et al.'s (2023) arguments about generalized causal claims, that such work should clearly state and then test the causal mechanisms for lack of replication/generalizability (you choose!), rather than simply state them as possibilities and move on (see also Bryan et al., 2021 for a discussion of this approach). Initiatives such as the Psychological Science Accelerator (Moshontz et al., 2018) and Many Labs (Ebersole et al., 2020) begin to move us in this direction. Certainly, increasing sample diversity is a necessary first step to being able to realize this vison. Increasing sample diversity during the discovery phase could lead to previously overlooked observations (Adetula et al., 2022; Majid, 2023; Syed, 2022), which has been illustrated in the recent GWAS literature (Abdellaoui et al., 2023; Saunders et al., 2022; Wojcik et al., 2019). Ironically, this approach is entirely consistent with a universal ontology, because the ultimate goal is to establish universally-applicable models, they are just ones that account for variability. What this does require, like my previous recommendation, is divestment from universal claims based on insufficient data.

This paper highlights the positive role of thinking about the diversity of our samples, and how considering sample diversity can improve our understanding of more "general" metascientific concepts. Although the examples largely focus on psychology, the arguments here have broad relevance across the sciences, many of which illustrate the dynamics laid out here.

References

Abdellaoui, A., Yengo, L., Verweij, K. J. H., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2022.12.011

Adetula, A., Forscher, P. S., Basnight-Brown, D., Azouaghe, S., & IJzerman, H. (2022). Psychology should generalize from—not just to—Africa. *Nature Reviews Psychology, 1*, 370-371. https://doi.org/10.1038/s44159-022-00070-y

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Baumeister, R. (2019). Self-control, ego depletion, and social psychology's replication crisis. *PsyArXiv*. https://doi.org/10.31234/osf.io/uf3cn

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*, 980-989. https://doi.org/10.1038/s41562-021-01143-3

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., & Laibson, D. I. (2015). The Fourth Law of Behavior Genetics. *Current Directions in Psychological Science*, *24*(4), 304–312. https://doi.org/10.1177/0963721415580430

Cheon, B. K., Melani, I., & Hong, Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science, 11*(7) 928-937. https://doi.org/10.1177/1948550620927269

Clarke, B., Schiavone, S., & Vazire, S. (2023). What limitations are reported in short articles in social and personality psychology? *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000458

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A Causal Framework for Cross-Cultural Generalizability. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/25152459221106366

Devezer, B., & Buzbas, E. (2021). *Minimum viable experiment to replicate* [Preprint]. http://philsci-archive.pitt.edu/21475/

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *ELife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Esterling, K., Brady, D., & Schwitzgebel, E. (2023). The necessity of construct and external validity for generalized causal claims. *OSF Preprints*. https://doi.org/10.31219/osf.io/2s8w5

Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., & Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, *28*(2), 243-250. https://doi.org/10.1038/s41591-021-01672-4

Graham, S. (1992). "Most of the subjects were White and middle class": Trends in published research on African Americans in selected APA journals, 1970–1989. *American Psychologist, 47*(5), 629. https://doi.org/10.1037/0003-066X.47.5.629

Guthrie, R. V. (1976). *Even the rat was white: A historical view of psychology.* Allyn & Bacon.

Haeffel, G. J., Burke, H., Vander Missen, M., & Brouder, L. (2023). What diverse samples can teach us about cognitive vulnerability to depression. *Collabra: Psychology*, *9*(1), 71346. https://doi.org/10.1525/collabra.71346

Hartmann, W. E., Kim, E. S., Kim, J. H. J., Nguyen, T. U., Wendt, D. C., Nagata, D. K., & Gone, J. P. (2013). In search of cultural diversity, revisited: Recent publication trends in cross-cultural and ethnic minority psychology. *Review of General Psychology, 17*(3), 243–254. https://doi.org/10.1037/a0032260

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42*(5), 443-455. https://doi.org/10.1037/0003-066X.42.5.443

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hewitt, J. K. (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics*, *42*(1), 1–2. https://doi.org/10.1007/s10519-011-9504-z

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences*, *113*(34), E4933–E4934. https://doi.org/10.1073/pnas.1608676113

Kahalon, R., Klein, V., Ksenofontov, I., Ullrich, J., & Wright, S. C. (2022). Mentioning the sample's country in the article's title leads to bias in research evaluation. *Social Psychological and Personality Science, 13*(2), 352–361. https://doi.org/10.1177/19485506211024036

Kullo, I. J., Lewis, C. M., Inouye, M., Martin, A. R., Ripatti, S., & Chatterjee, N. (2022). Polygenic scores in biomedical research. *Nature Reviews Genetics*, *23*(9), 524–532. https://doi.org/10.1038/s41576-022-00470-z

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) Evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*(4), 371–379. https://doi.org/10.1037/a0025172

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141

Majid, A. (2023). Establishing psychological universals. *Nature Reviews Psychology*, 1–2. https://doi.org/10.1038/s44159-023-00169-w

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, *100*(4), 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

Meehl, P. E. (1984). Foreword. In D. Faust (Ed.), *The limits of scientific reasoning (pp. 11–25).* Minneapolis: University of Minnesota Press

Melhuish, K., & Thanheiser, E. (2018). Research commentary: A rejoinder: Reframing replication studies as studies of generalizability: A response to critiques of the nature and necessity of replication. *Journal for Research in Mathematics Education*, *49*(1), 104–110. https://doi.org/10.5951/jresematheduc.49.1.0104

Moriguchi, Y. (2022). Beyond bias to Western participants, authors, and editors in developmental science. *Infant and Child Development*, *31*(1), e2256. https://doi.org/10.1002/icd.2256

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... & Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1*(4), 501-515. https://doi.org/10.1177/2515245918797607

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S. F. W., Oskarsson, S., Pickrell, J. K., Thom, K., Timshel, P., de Vlaming, R., Abdellaoui, A., Ahluwalia, T. S., Bacelis, J., Baumbach, C., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, *533*(7604), 539-542. https://doi.org/10.1038/nature17671

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D. A., Ahlskog, R., Magnusson, P. K. E., Oskarsson, S., Hayward, C., Campbell, A., … Young, A. I. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, *54*(4), 437-449. https://doi.org/10.1038/s41588-022-01016-z

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., Albrecht, E., Alizadeh, B. Z., Amin, N., Barnard, J., Baumeister, S. E., Benke, K. S., Bielak, L. F., Boatman, J. A., Boyle, P. A., … Koellinger, P. D. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, *340*(6139), 1467–1471. https://doi.org/10.1126/science.1235488

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science, 15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709

Roberts, S. O., & Mortenson, E. (2022). Challenging the White = Neutral Framework in Psychology. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916221077117

Saunders, G. R. B., Wang, X., Chen, F., Jang, S.-K., Liu, M., Wang, C., Gao, S., Jiang, Y., Khunsriraksakul, C., Otto, J. M., Addison, C., Akiyama, M., Albert, C. M., Aliev, F., Alonso, A., Arnett, D. K., Ashley-Koch, A. E., Ashrani, A. A., Barnes, K. C., … Vrieze, S. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, *612*(7941), 720-724. https://doi.org/10.1038/s41586-022-05477-4

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295. https://doi.org/10.1002/icd.2295

Sikorski, M., & Andreoletti, M. (2023). Epistemic functions of replicability in experimental sciences: defending the orthodox view. *Foundations of Science*. https://doi.org/10.1007/s10699-023-09901-4

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123-1128. https://doi.org/10.1177/1745691617708630

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, *12*(1), 118–130. https://doi.org/10.1177/1948550619900572

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*(6), 929–930. https://doi.org/10.1177/1745691616674460

Syed., M. (2021). Reproducibility, diversity, and the crisis of inference in psychology. *PsyArXiv*. https://psyarxiv.com/89buj/

Syed, M. (2022). Building causal knowledge in behavior genetics without racial/ethnic diversity will result in weak causal knowledge. *Behavioral and Brain Sciences*. https://psyarxiv.com/t5jm3/

Teo, T. (2009). Psychology without Caucasians. *Canadian Psychology/Psychologie Canadienne*, *50*(2), 91–97. https://doi.org/10.1037/a0014393

Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., & Hruschka, D. (2019). Generalizability is not optional: Insights from a cross-cultural study of social discounting. *Royal Society Open Science*, *6*(2), 181386. https://doi.org/10.1098/rsos.181386

The Turing Way Community. (2021). The Turing Way: A handbook for reproducible, ethical and collaborative research (1.0.1). *Zenodo*. https://doi.org/10.5281/zenodo.5671094

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*(4), 411-417. https://doi.org/10.1177/1745691617751884

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., … Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, *570*(7762), 514-518. https://doi.org/10.1038/s41586-019-1310-4

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proceedings of the National Academy of Sciences*, *120*(6), e2208863120. https://doi.org/10.1073/pnas.2208863120

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120. https://doi.org/10.1017/S0140525X17001972