Promise, peril, and perspective: Addressing concerns about reproducibility in social–personality psychology

Harry T. Reis⁎, Karisa Y. Lee

The University of Rochester, United States

ARTICLE INFO

Available online 1 April 2016

ABSTRACT

Current discussion about the evidentiary value of published research in social–personality psychology includes elements ranging in their premise. Some deride current practices as fundamentally flawed and call for extensive, perhaps even revolutionary, changes. Others are more circumspect, seeing the discipline as essentially healthy while acknowledging the need for evolution in several particulars about how research is conducted and reported. The articles in this special issue of JESP offer a variety of useful suggestions and recommendations in this latter regard. Our commentary provides an overarching perspective on these articles, suggesting a framework for considering their proposals in a way that stresses the promise of our science rather than its limitations.© 2016 Elsevier Inc. All rights reserved.

1. Introduction

In a comprehensive review, Stoeber and Otto (2006) summarized evidence about two kinds of perfectionism. The first, which they called perfectionistic strivings, concerns striving for high personal standards of performance, and is associated with, among various attributes, performance benefits, lower levels of self-doubt and self-criticism, and higher levels of conscientiousness. Other researchers have referred to this type of perfectionism as healthy perfectionism (e.g., Parker, 1997) and adaptive perfectionism (e.g., Ashby and Rice, 2002). The second form of perfectionism, which they labeled perfectionistic concerns, describes high levels of worry about mistakes and living up to others’ standards, doubts about the adequacy of one’s actions, and a perceived discrepancy between one’s achievements and expectations. This type of perfectionism, named by others as dysfunctional or maladaptive perfectionism, often impairs performance and is associated with neuroticism, self-doubts, and negative affect.

The current and explosive crisis about the reproducibility of findings in social–personality psychology often seems to straddle this distinction. We are of course not privy to the motives of those scholars who have contributed to the current discussion, but, it seems to us, evidence of both orientations can be readily discerned in the many articles and blog posts that have appeared on the subject. That is, some commentators seem focused on promoting high quality research by providing researchers with the necessary tools and incentives, while upgrading appropriate standard practices in the field. Others have written in a more apprehensive tone, emphasizing shortcomings in the field’s literature and modus operandus, advocating the need for field-wide higher standards and stricter enforcement of controls on the practices of individual researchers. More generally, these scholars seem focused on preventing mistakes and lamenting inadequacies in the knowledge base that social–personality science has produced (so far).

We believe that the field (and, for that matter, society as a whole) would be better served by modeling its guiding orientation and specific operational practices along the lines of adaptive perfectionism while eschewing those assumptions, principles, and rules that connote maladaptive perfectionism. In other words, we propose that the goal of achieving sound scientific insights and useful applications will be better facilitated over the long run by promoting good scientific practice rather than by stressing the need to prevent any and all mistakes. This proposal reflects a position one of us argued elsewhere (Finkel et al., 2015), called the error balance (EB) approach. The EB perspective argues that reducing false positives is only one strategy for improving the evidentiary value of our research. A broader perspective suggests that advancing the field requires a balance of the risks of false-positive and false-negative errors, acknowledging the simultaneous importance of empirical rigor, novel theories and insights, and inclusiveness of diverse topics and approaches.

This commentary does not review in turn each of the articles in this special issue of JESP. Rather, building on the distinction between perfectionistic strivings and perfectionistic concerns, we comment on four general themes that we see in these articles, themes that reflect many key points in the wider discussion that has been going on. These themes are: (1) improving research practice, (2) understanding how to think about replications, (3) genuinely attending to the differential demands of different research topics and methods, and (4) considering the wider implications of this debate. Our hope is that these comments
help move the field away from hand-wringing and toward striving for the best science that we can do.

2. Improving research practice

If there is anything on which all researchers might agree, it is the call for improving our research practices and techniques. Nevertheless, there are different ways of construing this call. Some might see it as a concern about perfection, or, to use Higgins's (1998) framing, in a prevention sense—preventing and correcting flaws that make our findings problematic. We see it more as a striving for perfection, in a promotion sense—scientists are never complacent with existing methods, because innovation and improvement fuel discovery. Greenwald (2012) offered an example that highlights this latter promotion approach. In describing the synergy between methodological innovation and theoretical advances, he noted that the majority of Nobel Prizes in the sciences have been awarded for the former: New methods evoke questions that could not have been investigated, much less conceived, with older (and presumably cruder) techniques. The resulting theoretical advances in turn suggest the need for further advances in methodology, instantiating a virtuous cycle that to us is plainly evident across the past century, roughly speaking, of social—personality psychology (Reis, 2010).

The contemporary literature indicates just how central methodological innovation has been to advancing the field. Any recent textbook includes many theories and findings generated by modern tools for assessing implicit cognition, neural activity, psychophysiology, and genetic influences on behavior, as well as methods for monitoring ambulatory activity and experience in naturalistic contexts. Additionally, Internet-based methods have greatly expanded opportunities for collecting data with large and diverse samples. Statistical methods have also advanced exponentially, to the point where we can routinely conduct pre-publication independent replications (PPIRs). This helps move the field toward the kind of research that can be conducted with minimal effort in relatively brief periods of time. PPIRs are not a reasonable ambition for many important topics in social—personality psychology for example, behavioral observations for improving our research practices and techniques. Nevertheless, as several of the authors in this special issue point out, not all replications are equally valuable. We concur with several authors (Crandall and Sherman, Stroebe) that conceptual replications offer the greatest potential to our field. We do research to corroborate theories, not the findings of individual studies. Validity is best established, as Brewer and Crano (2014) note, through programs of research, each study of which addresses a different concern (alternative explanations, method—boundariness, contextual boundaries, generalizability, and so on). Much of the current debate, however, is focused narrowly on direct or exact replications—whether the findings of a given study, carried out in a particular way with certain specific operations, would be repeated. Although exact replications are surely desirable, the papers by Fabrigar and by Crandall and Sherman remind us that in an absolute sense they are fundamentally impossible in social—personality psychology (see also Stroebe and Strack, 2014). Consequently, before conducting multiple “close” replications—a helpful terminological distinction articulated by Huffmeier—of individual studies, researchers would do well to explicitly establish the equivalence of “irrelevancies”—details of sampling, measurement, or procedure that in principle should have little or no influence on theoretically relevant variables (Cook, 1990). Irrelevancies are particularly critical in field replications, discussed by Maner, inasmuch as experimental control over them is necessarily weaker; Reis and Gosling, 2010. In this light, Fabrigar’s insistence that researchers take more care to demonstrate psychometric invariance is well-placed (see also Widaman and Grimm, 2014).

Of greater concern to us is Uhlmann’s suggestion that researchers routinely conduct pre-publication independent replications (PPIRs). Two issues are pertinent here. First, the suggestion that such replications may be carried out in laboratories that have little expertise implies that these studies may be conducted poorly, or at least with less-than-desirable levels of proficiency. Errors caused by low expertise or inadvertent changes are often catastrophic, in the sense of causing a study to fail completely, as Stroebe nicely illustrates. Unfortunately, these shortcomings rarely receive the level of publicity that the announcement of a “failure to replicate” does.

Second, and more problematically, incentivizing PPIRs, as Uhlmann proposes, would push the field toward the kind of research that can be conducted with minimal effort in relatively brief periods of time. PPIRs are not a reasonable ambition for many important topics in social—personality psychology—for example, behavioral observations of couple interactions, studies of group interaction and processes, longitudinal studies of health or personality, research that requires samples that are difficult to obtain, ambulatory assessments of naturalistic activity, field observations or interventions, and, for that matter, behavioral studies of the sort that were popular during social psychology’s so-called “golden age.” Rather, PPIRs seem suited to the sort of studies that Baumeister et al. (2007) caricatured as “the psychology of self-reports and finger movements.” Incentivizing researchers to move in this methodologically simplistic direction to us seems antithetical to

3. How to think about replications

Pointing to the importance of replication is a little like arguing that we should help the elderly and babies: everyone agrees in principle but how to make it happen can often become vexing. Still, if there was ever an uncontroversial statement with which all researchers could agree, it is that replications are valuable (Funder et al., 2014). Replication is, after all, the bedrock of science: For science to be a cumulative and self-correcting enterprise, as Crandall and Sherman indicate, replications be their results supportive, qualifying, or contradictory, must occur.

Nevertheless, as several of the authors in this special issue point out, not all replications are equally useful. We concur with several authors (Crandall and Sherman, Stroebe) that conceptual replications offer the greatest potential to our field. We do research to corroborate theories, not the findings of individual studies. Validity is best established, as Brewer and Crano (2014) note, through programs of research, each study of which addresses a different concern (alternative explanations, method—boundariness, contextual boundaries, generalizability, and so on). Much of the current debate, however, is focused narrowly on direct or exact replications—whether the findings of a given study, carried out in a particular way with certain specific operations, would be repeated. Although exact replications are surely desirable, the papers by Fabrigar and by Crandall and Sherman remind us that in an absolute sense they are fundamentally impossible in social—personality psychology (see also Stroebe and Strack, 2014). Consequently, before conducting multiple “close” replications—a helpful terminological distinction articulated by Huffmeier—of individual studies, researchers would do well to explicitly establish the equivalence of “irrelevancies”—details of sampling, measurement, or procedure that in principle should have little or no influence on theoretically relevant variables (Cook, 1990). Irrelevancies are particularly critical in field replications, discussed by Maner, inasmuch as experimental control over them is necessarily weaker; Reis and Gosling, 2010. In this light, Fabrigar’s insistence that researchers take more care to demonstrate psychometric invariance is well-placed (see also Widaman and Grimm, 2014).

Of greater concern to us is Uhlmann’s suggestion that researchers routinely conduct pre-publication independent replications (PPIRs). Two issues are pertinent here. First, the suggestion that such replications may be carried out in laboratories that have little expertise implies that these studies may be conducted poorly, or at least with less-than-desirable levels of proficiency. Errors caused by low expertise or inadvertent changes are often catastrophic, in the sense of causing a study to fail completely, as Stroebe nicely illustrates. Unfortunately, these shortcomings rarely receive the level of publicity that the announcement of a “failure to replicate” does.

Second, and more problematically, incentivizing PPIRs, as Uhlmann proposes, would push the field toward the kind of research that can be conducted with minimal effort in relatively brief periods of time. PPIRs are not a reasonable ambition for many important topics in social—personality psychology—for example, behavioral observations of couple interactions, studies of group interaction and processes, longitudinal studies of health or personality, research that requires samples that are difficult to obtain, ambulatory assessments of naturalistic activity, field observations or interventions, and, for that matter, behavioral studies of the sort that were popular during social psychology’s so-called “golden age.” Rather, PPIRs seem suited to the sort of studies that Baumeister et al. (2007) caricatured as “the psychology of self-reports and finger movements.” Incentivizing researchers to move in this methodologically simplistic direction to us seems antithetical to

1 Similar concerns have been voiced in other, putatively more objective disciplines. Bissell, a noted experimental cellular biologist, explained that “it is sometimes much easier not to replicate than to replicate studies, because the techniques and reagents are sophisticated, time-consuming, and difficult to master… Many scientists use epithelial cell lines that are exquisitely sensitive. The slightest shift in their microenvironment can alter the results—something a newcomer might not spot” (2013, p. 334). If anything, human participants seem far more likely than epithelial cells to be exquisitely sensitive to small (and likely undetected) shifts in the research environment.

2 One also wonders how tenure committees will evaluate the cases of individuals whose contributions to the field are of low expertise research.
developing a diverse, generalizable, and intellectually rich science. We elaborate on this point later.

A final comment about replication is to express our strong agreement with the viewpoint expressed by Fabrigar, Sakaluk, and Stroebe that when one views meta-analysis as the final arbiter of whether or not an effect is “real,” failed replications and underpowered studies may contribute to the literature and even strengthen a finding.3 Deeper consideration of the terms “failed” and “underpowered” may reveal just how limited the field is by dichotomous thinking. “Failed” implies that a result at p < .05 is somehow inferior to one at p > .05, a conclusion that scarcely merits disputation. (We are reminded of Rosnow and Rosenthal’s pithy quote, “Surely, God loves the .06 nearly as much as the .05,” 1989, p. 1277.) Similarly, .80 has been reified as the gold standard of power, a practice that would surprise and perhaps dismay its original proponent, who asserted that “(t)his arbitrary but reasonable...convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc” (Cohen, 1969, p. 54).

Moving away from dichotomous thinking about individual studies—a finding is significant or not, a study is adequately powered or not—and instead emphasizing trends across multiple studies (both in terms of p-values and effect sizes) will facilitate development of cumulative science.

4. Differential demands of different research topics

As noted earlier, many of the suggestions offered during the current crisis about reproducibility seem to us to be excellent ways to advance the field’s methodology. However, some of these proposals and attempts to formalize stricter empirical standards largely ignore the differential demands of different research domains and their associated methods, while implicitly tilting future research toward specific topics and methods. This topic has been addressed elsewhere (Finkel et al., 2015), specifically with regard to the feasibility of enacting six recommendations and requirements within one area of social–personality psychology, relationship science. These concerns apply to many other sub-disciplines within social–personality psychology, and to neighboring disciplines such as developmental and health psychology. Our goal here is not to reprise the Finkel et al. argument, however, we would like to briefly consider certain suggestions offered in this special section of JESP with the goal of highlighting how the intrinsic demands of our various research domains and methods call for deeper consideration.

For example, as discussed earlier, Uhlmann’s pre-publication independent replication (PPIR) approach is an attempt to improve the reproducibility of emerging research by replicating findings prior to their publication. However, like most existing types of replication, PPIRs are far more feasible for simple studies than for studies that involve lengthy or complex protocols. Thus, PPIRs also contribute to a problem created by other large-scale replication projects (e.g., Open Science Collaboration, 2012, 2015): the studies that are most often replicated are the ones that offer easy, relatively speedy execution. This can paint a skewed view of our science. For example, a recent p-curve analysis of the replications included in the Open Science Collaboration’s Reproducibility Project showed that the failures to replicate were largely limited to studies that presented methodological challenges (i.e., that had protocols that were considered difficult to carry out) and that provided opportunities for experimenter bias (Gangestad et al., 2016).4 Similarly, frameworks such as Sakaluk’s “exploring small, confirming big” approach have limited practicality for research domains involving intensive or time-consuming (e.g., longitudinal) research designs.

Although many of the current recommendations have undeniable value for certain research areas, we have reservations about treating any particular proposal as a rule, or even as a default (as Jussim and colleagues propose). While others have argued that the various open science guidelines should serve as suggestions and not requirements (e.g., Spellman, 2015), we worry that not everyone shares this view, and that refaying any of the various proposals as a “best practice” for research integrity may marginalize researchers and research areas that study phenomena or use methods that have a harder time meeting these requirements. For example, making available to interested parties data sets involving couples, families, or groups could compromise confidentiality, because individuals might be able to identify their partners’ responses by matching those responses with their own data. As another example, the 21-word solution proposed by Simmons et al. (2012) is unrealistic for large, complex multi-investigator, multi-method studies and data sets, for which protocols can run into the hundreds of pages and that may include hundreds of variables of interest to a diverse team of researchers. A third example is that preregistration may impair the process of disseminating exploratory findings by construing such work as inherently tentative. In our view, any general set of principles that label certain practices as better (i.e., more rigorous or trustworthy) than other practices must account for the intrinsic realities of different research domains and methods. If not, such principles, no matter how well-intentioned, invite the possibility of discrimination, not only within the field but also by decision-makers who are not privy to these realities.

Beyond the possibility of marginalizing certain subfields, it is important to consider how the adoption of new practices and standards may unintentionally lead the field to emphasize particular research topics and methods. For example, Baumeister et al. described how, after the cognitive revolution of the 1980s, researchers noticed “the success, prestige, and funding that could accrue to studies exploring inner psychological processes while postponing, perhaps indefinitely, examination of the behavioral results of these processes” (2007, p. 398). In our view, many of the new guidelines and practices that have been proposed, and in some cases implemented, similarly tip the scale toward short surveys and relatively simple experiments, adding to their attractiveness and thus pushing the field toward the kind of one-shot (“finger movement”) studies that Baumeister et al. (2007) derided as detrimental to scientific progress in social–personality psychology.

5. Considering broader implications

Undoubtedly, the current attention to reproducibility and research practices has brought and will continue to bring significant changes to the landscape of social–personality psychology.5 As these changes take hold and evolve, it is crucial that we consider some of their wider implications. First, in order for behavior to change, a legitimate cultural/institutional change must occur at the level of publication, hiring, funding, promotion, and tenure decisions. There is currently, as has often been noted, “a disconnect between what is good for scientists and what is good for science” (Nosek et al., 2012, p. 616). We cannot realistically expect researchers to change their practices when the implicit (and

3 Of course it is nonetheless heartening that eight out of ten of Uhlmann’s PPIR studies did replicate and the two that did not were originally null findings. This highlights an interesting question raised by Hufnemer et al. and by Stroebe: At what point do additional replications become unfruitful?
4 These authors note that it is impossible to determine whether experimenter effects biased results toward supporting the original positive result or against those results in the failed replication.
5 We disagree with Hales that the current crisis is more serious than the “social psychology as history” crisis of the 1970s. Although superficially that crisis centered on questions about whether phenomena generalized across time and settings, as Hales asserts, at a deeper level it implied profound doubts about whether social psychological theories were describing something real about human behavior or mere artifacts of time and place. As such, that crisis raised questions about the fundamental value and meaningfulness of our endeavor. We suggest that the ubiquity of the Internet and greater public attention to our work may be making the current debate seem more consequential.
perhaps even the explicit) standards remain the same. Recently, a colleague of ours relayed his frustrating experience of submitting a manuscript that included one null-result study among several studies with statistically significant findings. He was met with rejection after rejection, all the while being told that the null finding weakened the results or confused the manuscript. It is unreasonable to expect researchers to devote time and energy attempting to meet these new standards if institutional evaluation systems, both within the field and at our universities, are not yet there.

Another important consideration asks, at what point might the benefits of this debate outweigh its costs? We agree with Crandall and Sherman, and also Stroebe, that social psychology is, like all sciences, a self-correcting enterprise. Indeed, it might be argued that the current debate about research practices is in itself one example of such self-correction. But when does this debate stop facilitating progress and start becoming overly self-critical? Social psychologists are often reputed to be among the most severe critics of work within their own discipline. Shelley Taylor noted this tendency back in 1999:

“We are known to be one of the more self-destructive areas in psychology. We are legendary for trash[ing] each other’s work when we sit on grant review committees or act as external examiners. In our efforts to appear hard-nosed and smart, we inadvertently sabotage the work of our field. … If you’ve served on a review panel or been on an APA committee, you know that our reputation for supporting each other is poor. What you may not know is that it contributes to the low esteem in which our science is held. The sense of outsiders is that, if we do research that even we don’t respect, why should anyone else pay attention to it?” (p. 6).

Perhaps this inclination has not changed much since then. To the extent that the current commentary of crisis focuses on flaws and failures in existing research, and the consequent inadequacy of conclusions based on that work, it seems problematic to expect other scientists, funding agencies, students who may be considering a career in our field, and, for that matter, society as a whole, to take our work seriously. On the other hand, framing these developments in a more promotion-oriented manner, namely as an opportunity for improvement—“Gee, isn’t it great how our science is advancing? We already know a lot; with these practices, we can learn even more”—seems more likely to foster healthy respect for our field.

In this light, it seems pertinent to ask, what is the fundamental purpose of a scientific publication? Although it is often argued that journals are a “repository of the accumulated knowledge of a field” (APA Publication Manual, 2010, p. 9), equally important are their communicative functions. It seems to us that the former purpose emphasizes perfectionism (and especially perfectionistic concerns) whereas the latter purpose prioritizes sharing information so that others may build on a low esteem in which our science is held. The sense of outsiders is that, if we do research that even we don’t respect, why should anyone else pay attention to it?” (p. 6).

Perhaps this inclination has not changed much since then. To the extent that the current commentary of crisis focuses on flaws and failures in existing research, and the consequent inadequacy of conclusions based on that work, it seems problematic to expect other scientists, funding agencies, students who may be considering a career in our field, and, for that matter, society as a whole, to take our work seriously. On the other hand, framing these developments in a more promotion-oriented manner, namely as an opportunity for improvement—“Gee, isn’t it great how our science is advancing? We already know a lot; with these practices, we can learn even more”—seems more likely to foster healthy respect for our field.

In this light, it seems pertinent to ask, what is the fundamental purpose of a scientific publication? Although it is often argued that journals are a “repository of the accumulated knowledge of a field” (APA Publication Manual, 2010, p. 9), equally important are their communicative functions. It seems to us that the former purpose emphasizes perfectionism (and especially perfectionistic concerns) whereas the latter purpose prioritizes sharing information so that others may build on a body of work and ideas (flawed as they may eventually turn out to be), Thus, while formal conceptual reasoning clearly has an important place in theory development, as Schaller notes, we also believe that it is often acceptable to do research just to see what will happen. Informed curiosity can motivate research that provides valuable observations and insights (Rozin, 2001). As Rozin further explained, consistent with our earlier point that validity is a property of programs of study, “the best hope we may have… is to accumulate flawed (ambiguous) evidence in large amounts and from many different sources and approaches. This is probably the only practical route to understanding Homo sapiens in a social context” (2001, p. 3). Perfectionistic striving advances knowledge and can lead to important applications of research, but we should not get so caught up in perfectionistic concerns that they impede the rapid accumulation and dissemination of research findings.

6. Conclusion

We began this commentary by describing a difference between two kinds of perfectionism, one motivated by high personal standards and the desire to improve and the other motivated by self-doubts and a pervasive sense of inadequacy. We see signs of both tendencies in the current discussion. Many of the proposals in this collection of articles, as well as many other recommendations described elsewhere, seem like worthwhile examples of the former. On the other hand, many other examples seem to us to be instances of the latter. Our commentary describes several themes that may help researchers and other interested parties distinguish between more and less adaptive proposals for advancing the science of social–personality psychology.

Enhancing the evidentiary value of our research is surely a goal all social–personality scientists share (Finkel et al., 2015). As debate continues about the relative merit of each proposal for accomplishing this goal, we might be mindful about the ultimate purpose of higher standards. Moving our science forward is a pursuit that requires thoughtful and selective balance among several elements inherent in any progress we might make—rigorously testing and establishing our theories while energizing the discovery process, promoting careful yet innovative applications of our findings and insights for the common good, utilizing existing resources effectively and efficiently, and fostering public acceptance and support of the field. New standards and practices are advances only if they facilitate rather than hinder the sum total of all these elements.

References
