https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

Commentaries

Commentaries and Rejoinder on Klein et al. (2014)

The Limits of Direct Replications and the Virtues of Stimulus Sampling

Benoît Monin[1] and Daniel M. Oppenheimer[2]

[1]Stanford University, Stanford, CA, USA, [2]University of California, Los Angeles, CA, USA

Abstract. While direct replications such as the "Many Labs" project are extremely valuable in testing the reliability of published findings across laboratories, they reflect the common reliance in psychology on single vignettes or stimuli, which limits the scope of the conclusions that can be reached. New experimental tools and statistical techniques make it easier to routinely sample stimuli, and to appropriately treat them as random factors. We encourage researchers to get into the habit of including multiple versions of the content (e.g., stimuli or vignettes) in their designs, to increase confidence in cross-stimulus generalization and to yield more realistic estimates of effect size. We call on editors to be aware of the challenges inherent in such stimulus sampling, to expect and tolerate unexplained variability in observed effect size between stimuli, and to encourage stimulus sampling instead of the deceptively cleaner picture offered by the current reliance on single stimuli.

Keywords: Stimulus sampling, direct replication, mixed models, random effects

Klein et al.'s (2014) replication project renders a tremendous service to psychology, countering recent Chicken Little catastrophism by demonstrating that many published effects are quite reliable. We also thank Klein et al. for bringing an obscure paper of ours to the attention of new readers. Despite the successful replication of our effect, we appreciate the opportunity to comment. We agree that direct replications are important. Simmons, Nelson, and Simonsohn (2011) rightly criticized so-called "conceptual replications" because variations in procedure enable the original authors to dismiss null results too easily. But we want to point to the limits of direct replication in determining the robustness of an effect beyond its habitat of original discovery.

Klein et al. chose to replicate two conditions from the three originally in our Study 2a (Oppenheimer & Monin, 2009): Participants imagined walking into a room and seeing a man roll three dice which came up (depending on condition) (6,6,6) or (6,6,3), and estimated how many times he had rolled prior to being observed. The effect, which replicated so well, was that individuals thought that the man who rolled the triple 6 had been throwing dice for longer. Given 30 significant replications out of 36 attempts, we can be fairly confident that a man throwing (6,6,6) seems like he's been throwing for longer than a man throwing (6,6,3). The fact that the laboratories did not vary the sequence [e.g., to (6,6,5), or (3,3,3)], or the context [e.g., to a lottery, a horse race, or an everyday chance situation] has obvious virtue in facilitating cross-sample comparison. But such direct replications do little to increase our confidence that an observed effect generalizes to other situations, and yield estimates of effect size that may be atypical.

Judd, Westfall, and Kenny (2012) recently pointed to a "pervasive but largely ignored problem": statistical analyses in psychology routinely ignore that stimuli are sampled from a larger population. An equally pervasive and ignored problem is that stimuli are frequently not sampled at all, and instead a single vignette or a single social target is routinely assumed to stand for a larger class. Psychologists understand the need to sample subjects to generalize across people, and the present paper highlights the benefits of sampling across laboratories. But stimulus sampling is still largely neglected in psychology (see also Wells, & Windschitl, 1999).

Of the 12 other studies replicated by Klein et al. (2014), only one (Jacowitz & Kahneman, 1995) explicitly includes stimulus sampling with 15 topics (4 retained in the replication). Besides the two IAT studies (which do sample stimuli

© 2014 Hogrefe Publishing

DOI: 10.1027/1864-9335/a000202

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

but do not statistically treat them as random factors), the remaining nine effects all rely on a single vignette or stimulus set; six rely on a single-item response to a single vignette. Note that including several scale items (as in Caruso, Vohs, Baxter, & Waytz, 2013) or a unique set of multiple stimuli (as in Carter, Ferguson, & Hassin, 2011) does not qualify as stimulus sampling if the analysis focuses on one aggregate response – obscuring whether effects may result from a single item or stimulus, and preventing the estimation of variability in effect size between stimuli.

We would not have found our effect convincing on its own, however many laboratories replicated it, if it had only relied on the manipulation of a single vignette. We must credit our *JDM* reviewers for suggesting that we sample over situations, and our editor, Jonathan Baron, for suggesting that we use a mixed-model approach to analyze those. The result is that our Study 3 included 16 different situations involving repeated chance events. The effect was stronger for some stimuli than others, and was not observed for some – but crucially, it emerged overall. Compared to the two-vignette demonstration that Klein et al. (2014) replicated across 36 laboratories, this 16-vignette demonstration in 1 laboratory increased our confidence that the effect is more than an artifact of a specific scenario.

Unfortunately, the incentive structures that impede direct replications similarly undermine attempts to improve stimulus sampling. Consider a researcher who increases the number of vignettes from 1 to 4. While this is undeniable progress over the single-vignette approach, 4 is not enough to treat vignette as a random factor (Bates, 2010), so she would have to include it as a fixed factor. This means that she might have to deal with interaction effects of no theoretical interest, and with reviewers demanding an explanation of why the effects are stronger with some vignettes than others. Thus, researchers who cannot afford to develop a large number of stimuli (e.g., 30) are incentivized to use only one stimulus or risk facing difficulty publishing. Journal editors could contribute to improving stimulus sampling norms by expecting and tolerating unexplained variability in small stimulus sets, especially in areas where the cost of stimulus generation is not trivial.

In conclusion, we applaud Klein et al.'s direct replication approach, but we also want to stress the need for better stimulus sampling, and worry that a narrow emphasis on direct replication can obscure the importance of multiple stimuli, vignettes, and diverse experimental paradigms. Direct replications show that specific effects can be generalized across subjects and laboratories, but ignore the issue of generalizing across contexts and stimuli. The solution entails a fundamental change in the way psychologists design their studies and analyze their data. Online studies make it easier to rotate multiple stimuli or vignettes over large samples, while new, free, open-source software (e.g., the lme4 package in R; Bates, Maechler, Bolker, & Walker, 2014) make it relatively easy for researchers to treat stimuli as random effects. We hope to see future replication projects, reflecting the field's changing practices, adopt a "Many Stimuli" approach, enabling us to build our confidence and appreciate variations in replicability not only across laboratories, but also across many versions of the constructs of interest.

Acknowledgments

The authors declare no conflict of interest with the content of this article. Contributor's statement: B.M. drafted this manuscript. B.M. and D.O. both engaged in substantial revision of the manuscript. The 2nd author is partially funded by NSF 1128786.

References

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6 [Computer software]*. Retrieved from http://CRAN.R-project.org/package=lme4

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science, 22*, 1011–1018.

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General, 142*, 301–306.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161–1166.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54.

Klein, R. A., Ratliff, R. A., Vianello, M., Adams, R. A. Jr., Bahník, S., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45*, 142–152. doi: 10.1027/1864-9335/a000178

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making, 4*, 326–334.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115–1125.

Published online May 30, 2014

Benoît Monin

Stanford Graduate School of Business
Knight Management Center
655 Knight Way
Stanford, CA 94305
USA
E-mail monin@stanford.edu

 © 2014 Hogrefe Publishing

Commentary on the Attempt to Replicate the Effect of the American Flag on Increased Republican Attitudes

Melissa J. Ferguson,[1] Travis J. Carter,[2] and Ran R. Hassin[2]

[1]Cornell University, Ithaca, NY, USA, [2]Colby College, Waterville, WE, USA, [3]Hebrew University, Jerusalem, Israel

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

Abstract. In this commentary, we reflect on what we have learned from the experience of being part of the ManyLabs replication, both in terms of the phenomenon being studied, and in terms of the mechanics of such an ambitious replication project. Replication attempts like this one are clearly valuable and will continue to inform our science. We discuss a number of lessons we have taken from the process related to the divide between direct and conceptual replication, and whether the data can inform our current theory regarding the original effect. In discussing these issues, we advocate for transparent flexibility in data analyses and the involvement of the original authors at every stage of the process.

Keywords: replication, priming, flag priming

We begin with two points. First, we thank the authors of the ManyLabs project (Klein et al., 2014) and believe that this foray in large-scale replication is important. Second, we do not dispute the conclusion that the result of Carter, Ferguson, and Hassin's (2011) Study 2 was not replicated.

Many in the field are identifying practices to improve our science, and an increasing willingness to conduct and publish replications can only help. This is a learning process, however, and we view the ManyLabs project not only as a replication of experiments, but also as an experiment in replications – one that should inform best practices. Here we make a few observations.

Direct Replication

A direct replication attempts to mimic the experimental methodology used in the original study. Some prefer direct over conceptual replications, which instead aim to extend a finding by using different procedures (e.g., Pashler & Harris, 2012; Simons, 2014). The authors of the ManyLabs project worked with us to ensure that their materials were nearly identical to ours. There are some aspects, however, that are difficult to mimic, making the distinction between "direct" and "conceptual" replications fuzzy.

For instance, national flags, ex hypothesis, activate knowledge related to one's nation that is shaped by the prevailing political atmosphere, which is hardly inert. The original experiment was run in 2009 – shortly after the first African-American was sworn in as president of the US – whereas the ManyLabs was run 4 years later, in the 5th year of his presidency. Knowledge (e.g., about political parties) associated with America changed over that time (e.g., Devos & Ma, 2012 and Ma & Devos, 2013 show that automatic associations with Barack Obama changed over this time). Although many effects should remain stable over time (e.g., numerical anchoring effects), stimuli that represent time-sensitive knowledge and events (e.g., political and national symbols) should be expected to change over time, making "direct" replications difficult (McGuire, 2013).

Consider also differences in samples. The ManyLabs authors do this by including contextual variables (e.g., lab vs. online) in their analyses. Although we did not have any a priori reason to expect in-lab versus online differences, there are significant differences between online and in-lab samples on nearly *all* the variables related to the replication of our study. This points to the necessity of including potential interactions with sample characteristics in plans and analyses to account for such differences.

Lastly, whereas our original study was a stand-alone experiment, participants in the ManyLabs study engaged in many experiments, many of which contained direct references to the US (i.e., 2, 3, 6, 7, and 10). This introduces variability in the manipulation of whether participants were primed (or not) by the US that cannot be accounted for by controlling for position. Although one could test those who completed the flag study first (which Klein et al. did), this results in a smaller sample size, making a test of our current theoretical model with moderators included (see below) underpowered (McClelland & Judd, 1993).

A different political atmosphere, different subject pools, and different states of mind separate the original and the replication attempt. For these reasons, we view this as a conceptual, and not a direct, replication. We can learn from its failure, just as we can learn from other recent studies that identify moderators to the original phenomenon (e.g., Kalmoe & Gross, 2013).

What About Theory?

Since the original study, we developed a theory to account for the dynamic nature of primes that depend on shifting cultural knowledge. We are testing how to identify those who

© 2014 Hogrefe Publishing

possess the relevant implicit associations, and are therefore most likely to be influenced by a flag prime (and how). The ManyLabs team graciously included our proposed moderators. Unfortunately, a test of our model is not possible because the contextual variables cannot be included in the analyses given the effects of those variables on our factors, as noted. The original effect was not replicated, but these data do little to confirm or disconfirm our current model.

The point is that replications are theory-laden. It may take time to develop theories that can fully account for the conditions required to observe a phenomenon (see Cesario, 2014). Hence, understanding the theory behind the effects – and contacting authors for their latest theoretical developments – is an important step in the process.

Data Analyses

In order to create a confirmatory design, the ManyLabs authors preregistered the analyses. This prevents post hoc hypotheses and befits replications. However, data can sometimes surprise us in ways that render the original plan insufficient.

For example, the use of hierarchical regression without including lower-order interaction terms can lead to misleading results. The ManyLabs analyses show that the two 3-way interactions we predicted based on our current model are at $p = .05$, and $p = .07$. However, when the predictors are first centered/standardized, then these interactions are significant at $p < .001$. To be clear, we conducted that these p values are misleading, and *do not* in fact reflect support for those interactions; when influential lower-order interactions are not included in a model, the higher-order interactions can be difficult to interpret. (We note that although Klein et al. conduct analyses to test our proposed moderators while including all lower order terms, these analyses still do not include the contextual variable (i.e., lab/online) and do not account for whether the sample could have been contaminated by previous ManyLabs studies mentioning the U.S.).

So, although pre-approved plans have their advantages, they should leave room for flexibility. There must be transparent ways to conduct additional analyses when warranted by the data themselves.

Author Contact and Peer Review

Beyond contacting the original authors for materials, procedures, and theoretical updates, it is important that replications are peer-reviewed in a way that will allow the discovery of unintentional flaws. Including the original authors as reviewers serves this goal, and an impartial editor can adjudicate the legitimate concerns while minimizing any motivated cognition. If fact, this is what they do daily.

For example, the figure of effect sizes that has become the symbol of this research includes international samples. Our original study tested the influence of an American flag on Americans' support for American political policies. There is nothing in the paper to suggest that this effect could be transplanted to other countries.

The bottom line? We believe the failure to conceptually replicate our original study will be informative. We also believe that this experiment in replicating teaches us about the do's and do not's of our future science. We thank again the ManyLabs authors.

Acknowledgments

The authors declare no conflict-of-interest with the content of this article. Author contributions: Analyzed data: M.F. and T.C.; Wrote paper: M.F., T.C., and R.H.

References

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science, 22*, 1011–1018. doi: 10.1177/0956797611414726

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48. doi: 10.1177/1745691613513470

Devos, T., & Ma, D. S. (2012). How ''American'' is Barack Obama? The role of national identity in a historic bid for the white house. *Journal of Applied Social Psychology, 43*, 214–226. doi: 10.1111/jasp.12069

Kalmoe, N. P., & Gross, K. (2013). *Priming America: Experimental Tests of Flag Imagery Effects in Presidential Elections*. Prepared for presentation at the 2013 Annual Meeting of the American Political Science Association, Chicago, IL, August 29 – September 1, 2013.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A ''many labs'' replication project. *Social Psychology, 45*, 142–152. doi: 10.1027/1864-9335/a000178

Ma, D. S., & Devos, T. (2013). *Every heart beats true, for the red, white, and blue: National identity predicts voter support*. *Analyses of Social Issues and Public Policy*. Advance Online Publication. doi: 10.1111/asap.12025

McClelland, G.H., & Judd, C.M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*(2), 376–390.

McGuire, W. J. (2013). An additional future for psychological science. *Perspectives on Psychological Science, 8*, 414–423. doi: 10.1177/1745691613491270

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536. doi: 10.1177/1745691612463401

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76–80. doi: 10.1177/1745691613514755

Published online May 30, 2014

Melissa J. Ferguson

Dept. of Psychology
230 Uris Hall
Cornell University
Ithaca, NY 14853-7601
USA
E-mail MJF44@cornell.edu

© 2014 Hogrefe Publishing

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

Support for the Replicability of Imagined Contact Effects

Richard J. Crisp,[1] Eleanor Miles,[2] and Shenel Husnu[3]

[1]University of Sheffield, UK, [2]University of Sussex, UK, [3]Eastern Mediterranean University, Cyprus

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

Abstract. As part of their Many Labs project Klein et al. (2014) replicated the effects of an imagined contact study carried out by Husnu and Crisp (2010). In their report the authors argue the data provides weak support for replicability. However, the effect observed was both significant and comparable to that obtained from a recent meta-analysis for the relevant outgroup. This suggests that the Many Labs project may provide stronger support for the existence of imagined contact effects than currently thought. We discuss the value in interpreting replications within the context of the existing literature.

Keywords: imagery, prejudice, imagined contact, interventions

Imagined intergroup contact (Crisp & Turner, 2009) is a new indirect contact strategy for promoting tolerance and more positive intergroup relations. As part of their Many Labs project Klein et al. (2014) replicated an imagined contact study originally carried out by Husnu and Crisp (2010). They concluded that this test revealed weak support for replicability but we suggest that further consideration of their methods and data, particularly in the context of the wider literature, might serve to moderate this conclusion.

Many Labs is not a Direct Replication

A recent meta-analysis of over 70 imagined contact studies (Miles & Crisp, 2014) revealed a robust moderate effect of imagined contact on a range of dependent variables (attitudes, emotions, intentions, behavior) and toward a range of different groups (based on ethnicity, age, religion, etc., overall $d_+ = 0.35$). Klein et al. (2014) attempted to replicate one of these imagined contact studies originally carried out by Husnu and Crisp (2010). However, in the process of implementing the Many Labs methodology, changes to original study procedures were required to adapt the original studies to the Many Labs template. In the case of Husnu and Crisp (2010) simplification of the procedure resulted in a lack of specificity regarding participants, which renders the replication effect difficult to interpret. This is because the basic imagined contact effect is that imagining contact with an outgroup member reduces prejudice toward that outgroup. As participants in the Many Labs study were not asked to report their religion, the ingroup versus outgroup status of the participants relative to the imagined targets is unknown. Given the wide net cast in the Many Labs sample it is likely that a subset of participants imagined contact with, and reported prejudice toward, their own ingroup. This could undermine the usefulness of comparisons made with the original study, and arguably limits the conclusions that can be drawn about the underlying effect.

More generally, the Many Labs project did not consider the interaction between country context and outgroup identity. The original Husnu and Crisp study examined effects on prejudice when British students imagined contact with British Muslims. In line with the notion that ingroup and outgroup identity are key to the imagined contact effect, Klein et al. (2014) modified the Muslim outgroup for their Turkish sample, who instead imagined contact with Christians. However, this modification highlights the complexity of cultural context, as Christians are far less likely to be considered as an outgroup by Turkish participants than Kurds or Armenians, who are ethnic minorities with a long history of conflict within Turkey (Bikmen & Sunar, 2013). Of course, the potential for variability in imagined contact effects as a function of culture and context is entirely consistent with what we already know from cross-cultural psychology: There are huge cultural differences in the meaning, status, and relations between what are, on the face of it, the same outgroups. Thus, the tests constituting the Many Labs replication likely encompass numerous moderators of the effect, and it is likely the single overall effect reported in the Many Labs paper masks important differences among these multiple tests. In sum, while undoubtedly valuable, we would argue that the Many Labs study is not a direct replication of the original Husnu and Crisp study. Rather, it is an important new data point in efforts to refine our understanding of the effect relating to different cultural groups, in different cultural contexts.

Many Labs Observes an Effect Size Consistent with Meta-Analytic Estimates

Klein et al. (2014) draw conclusions not only about the replicability of the original study, but about the existence of the underlying imagined contact effect. Specifically, although they obtained a significant effect, because it was

© 2014 Hogrefe Publishing

smaller that that obtained in the original study, they argue that their findings constitute weak support for the replicability of imagined contact effects. However, this conclusion does not take into account other relevant evidence concerning the existence and magnitude of these effects – particularly meta-analytic evidence.

To answer the question of whether replication attempts provide evidence for the existence of the underlying effect, the most meaningful point for comparison is arguably not the original effect size, but the effect size from a meta-analysis. Miles and Crisp's (2014) meta-analysis of over 70 imagined contact conceptual replications has already established that the true overall effect size for imagined contact is lower than that observed by Husnu and Crisp (2010; 0.86). In fact, the refined estimate for the entire sample is 0.35, with our estimate for religious groups (those tested in the current replication attempt) being 0.22. Thus the observed effect size of 0.13 in the Many Labs study is substantively different from the original Husnu and Crisp study, and from our overall estimate of 0.35, but *not* from the most appropriate comparison: The meta-analytic estimate for religious outgroups (0.22). In other words, this is a replication effect size consistent with previous meta-analytic estimates of the effect size for the relevant outgroup.

Conclusion

The investigations offered by the Many Labs paper provide important insights in to the replicability of key psychological findings. However, we would not want to draw conclusions about a general effect from a single new study, no matter how large. In the case of Husnu and Crisp (2010), a wider view reveals that the observed effect size is consistent with the accumulated evidence for the effect of imagining contact with religious outgroups. We therefore suggest the current data provides converging and qualified support for imagined contact effects, while providing a valuable template for continuing efforts to refine, clarify and define the extent and applicability of the technique.

Acknowledgments

The authors declare no conflict of interest with the content of this article. The authors made the following contributions to this paper: Wrote paper: R.C., E.M., S.H.

References

Bikmen, N., & Sunar, D. (2013). Difficult dialogs: Majority group members' willingness to talk about inequality with different minority groups. *International Journal of Intercultural Relations, 37*, 467–476.

Crisp, R. J., & Turner, R. N. (2009). Can imagined interactions produce positive perceptions? Reducing prejudice through simulated social contact. *American Psychologist, 64*, 231–240. doi: 10.1037/a0014718

Husnu, S., & Crisp, R. J. (2010). Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology, 46*, 943–950. doi: 10.1016/j.jesp.2010.05.014

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45*, 142–152.

Miles, E., & Crisp, R. J. (2014). A meta-analytic test of the imagined contact hypothesis. *Group Processes and Intergroup Relations, 17*, 3–26. doi: 10.1177/1368430213510573

Published online May 30, 2014

Richard J. Crisp

Department of Psychology
University of Sheffield
Western Bank, Sheffield, S10 2TP
United Kingdom
E-mail r.crisp@sheffield.ac.uk

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

 © 2014 Hogrefe Publishing

Does Merely Going Through the Same Moves Make for a "Direct" Replication?

Concepts, Contexts, and Operationalizations

Norbert Schwarz[1] and Fritz Strack[2]

[1]University of Southern California, Los Angeles, CA, USA, [2]Universität Würzburg, Germany

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

Abstract. Klein and colleagues (2014) conducted a "direct" replication of a study on the influence of frequency scales on behavioral reports. To do so, they administered a scale based on behavioral frequencies in a 1983 German sample to diverse samples whose (known) 2013 behavioral frequencies exceeded the historical values by a factor of two, resulting in a replication that is "technically equivalent" while missing the realization of psychologically equivalent differences between behavior and response scale. We discuss the difference between technical and psychological equivalence and highlight its implications for testing the robustness of psychological phenomena.

Keywords: replication, operationalization, methodology

The "Many Labs" replication project (Klein et al., 2014) included a replication of an experiment on the influence of response scales on behavioral reports (Schwarz, Hippler, Deutsch, & Strack, 1985). Consistent with several dozen prior experiments (for a partial review, see Schwarz, 1999), the original finding was robust: people report higher behavioral frequencies along scales that present high rather than low response alternatives. This reflects that memory for mundane behaviors is poor, which forces people to estimate what their absolute behavioral frequency might be. To do so they draw on the range of scale values as a frame of reference, which is licensed by the assumption that researchers construct meaningful scales that reflect the actual distribution of interest (Schwarz, 1994). Accordingly, the observed effects increase with variables that increase the need to estimate, and decrease with variables that call the applicability and relevance of the scale into question (for a review, see Schwarz, 1999). Put simply, the processes underlying scale effects are context sensitive.

One important implication of context sensitive processes is that merely going through the same technical moves does not amount to equivalent tests of the same process when the context changes. In the present case, the original study, conducted in Germany in 1983, was designed for a population with an average TV consumption of slightly more than two hours a day (Darschin & Frank, 1982). To ensure comparable plausibility of both scale versions, the average TV consumption served as the high end of the low frequency scale (running from "up to ½ hr" to "more than 2½ hr") and the low end of the high frequency scale (running from "up to 2½ hr" to "more than 4½ hr"). This procedure provides a straightforward recipe for a replication: construct a scale that covers equal ranges below and above the actual behavioral frequency of the population used in the replication. Based on Nielsen data (Marketing Charts, 2013) the average TV consumption in the United States was just over 5 hr/day in early 2013. Accordingly, both of the 1983 German scales present values below the likely behavior of the majority of a US sample, resulting in two differentially "low frequency" scales with no "high frequency" scale included in the replication; the match between the scale values and the behavior of the international samples included in the replications is unknown.

When Klein and colleagues requested the original materials, we suggested that the replication should follow the construction principles of the 1983 German scales rather than its historically bound specific numeric values (email exchange with Richard Klein, May 2013); we also noted that a more informative study would follow a 2 (high vs. low scale values) × 2 (based on 1983 German consumption vs. 2013 sample-adequate consumption)-design, for which the N of several planned samples would have been sufficient. In the interest of a "direct" replication, the authors chose to go with the historical German values, resulting in a replication that can be described as "technically direct" while missing the goal of realizing psychological conditions that are comparable to the original study. The same oversight impairs comparisons across the different samples used in the multiple replications – differential discrepancies between the scale value and the respective sample's behavior result in differential treatments. This prioritization of technical over conceptual equivalence threatens the key goal of the "Many Labs" project, which explicitly focuses on the robustness of effect sizes. Any meaningful comparison of effect sizes across studies has to ensure the psychological equivalence of the treatment rather than its mere technical equivalence. In the present case, the observed variation in effect sizes may reflect the variables that motivated the "Many Labs" project and/or differential discrepancies between the 1983 German scale values and

© 2014 Hogrefe Publishing

the actual behavioral frequencies in the samples used, which the authors decided to ignore.

In general, meaningful replications need to realize the psychological conditions of the original study. The easier option of merely running through technically identical procedures implies the assumption that psychological processes are context insensitive and independent of social, cultural, and historical differences (Cesario, 2014; Stroebe & Strack, 2014). Few social (let alone cross-cultural) psychologists would be willing to endorse this assumption with a straight face. If so, mere procedural equivalence is an insufficient criterion for assessing the quality of a replication. Instead, replications need to be evaluated within the theoretical framework under study, paying close attention to the extent to which technically identical procedures are, or are not, conceptually equivalent in a different context. This implies that the evaluation of replications cannot be limited to reviews of technical equivalence at the registration stage. Instead, meaningful evaluations require manipulation checks and other information that bears on treatment equivalence (e.g., the relationship between scale values and population behavior in the above studies), which usually requires that relevant data are available. We therefore noted with surprise that Nosek and Lakens' (2014) editorial discourages such evaluations as "CARKing" (critiquing after the results are known) and that the contributions to this special issue did not undergo post-registration review. Unfortunately, testing the robustness of a psychological phenomenon requires a theoretically informed analysis that goes beyond distributing copies of the original questionnaire while ignoring changes in context. Hence, replication reports need to make a persuasive case for the equivalence of treatment and appropriateness of interpretation, which need to be subjected to the same review process as any other scientific publication.

Acknowledgments

The authors declare no conflict of interest.

References

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48.

Darschin, W., & Frank, B. (1982). Tendenzen im Zuschauerverhalten [Trends in viewer behavior]. *Media Perspektiven, 4*, 276–284.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., …, & Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45*, 142–152. doi: 10.1027/1864-9335/a000178

Marketing Charts. (2013). *Are young people watching less TV?* Retrieved from http://www.marketingcharts.com/wp/television/are-young-people-watching-less-tv-24817/

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141. doi: 10.1027/1864-9335/a000192

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology, 26*, 123–162.

Schwarz, N. (1999). Frequency reports of physical symptoms and health behaviors: How the questionnaire determines the results. In D. C. Park, R. W. Morrell, & K. Shifren (Eds.), *Processing medical information in aging patients: Cognitive and human factors perspectives* (pp. 93–108). Mahwah, NJ: Erlbaum.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59–71.

Published online May 30, 2014

Norbert Schwarz

Department of Psychology
University of Southern California
3620 S. McClintock Ave
Los Angeles, CA 90089-1061
USA
E-mail norbert.schwarz@usc.edu

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

 © 2014 Hogrefe Publishing

Theory Building Through Replication
Response to Commentaries on the "Many Labs" Replication Project

Richard A. Klein,[1] Kate A. Ratliff,[1] Michelangelo Vianello,[2] Reginald B. Adams Jr.,[3] Štěpán Bahník,[4] Michael J. Bernstein,[5] Konrad Bocian,[6] Mark J. Brandt,[7] Beach Brooks,[1] Claudia Chloe Brumbaugh,[8] Zeynep Cemalcilar,[9] Jesse Chandler,[10,37] Winnee Cheong,[11] William E. Davis,[12] Thierry Devos,[13] Matthew Eisner,[39] Natalia Frankowska,[6] David Furrow,[15] Elisa Maria Galliani,[2] Fred Hasselman,[16,38] Joshua A. Hicks,[12] James F. Hovermale,[17] S. Jane Hunt,[18] Jeffrey R. Huntsinger,[19] Hans IJzerman,[7] Melissa-Sue John,[20] Jennifer A. Joy-Gaba,[17] Heather Barry Kappes,[21] Lacy E. Krueger,[18] Jaime Kurtz,[22] Carmel A. Levitan,[23] Robyn K. Mallett,[19] Wendy L. Morris,[24] Anthony J. Nelson,[3] Jason A. Nier,[25] Grant Packard,[26] Ronaldo Pilati,[27] Abraham M. Rutchick,[28] Kathleen Schmidt,[29] Jeanine L. Skorinko,[20] Robert Smith,[14] Troy G. Steiner,[3] Justin Storbeck,[8] Lyn M. Van Swol,[30] Donna Thompson,[15] A. E. van 't Veer,[7,31] Leigh Ann Vaughn,[32] Marek Vranka,[33] Aaron L. Wichman,[34] Julie A. Woodzicka,[35] and Brian A. Nosek[29,36]

[1]University of Florida, Gainesville, FL, USA, [2]University of Padua, Italy, [3]The Pennsylvania State University, University Park, PA, USA, [4]University of Würzburg, Germany, [5]Pennsylvania State University Abington, Abington, PA, USA, [6]University of Social Sciences and Humanities Campus Sopot, Sopot, Poland, [7]Tilburg University, Tilburg, The Netherlands, [8]City University of New York, New York, USA, [9]Koç University, Istanbul, Turkey, [10]University of Michigan, Ann Arbor, MI, USA, [11]HELP University, Kuala Lumpur, Malaysia, [12]Texas A&M University, College Station, TX, USA, [13]San Diego State University, San Diego, CA, USA, [14]Ohio State University, Columbus, OH, USA, [15]Mount Saint Vincent University, Nova Scotia, Canada, [16]Radboud University Nijmegen, The Netherlands, [17]Virginia Commonwealth University, Richmond, VA, USA, [18]Texas A&M University-Commerce, Commerce, TX, USA, [19]Loyola University Chicago, IL, USA, [20]Worcester Polytechnic Institute, Worcester, MA, USA, [21]London School of Economics and Political Science, London, UK, [22]James Madison University, Harrisonburg, VA, USA, [23]Occidental College, Los Angeles, CA, USA, [24]McDaniel College, Westminster, MD, USA, [25]Connecticut College, New London, CT, USA, [26]Wilfrid Laurier University, Waterloo, ON, Canada, [27]University of Brasilia, DF, Brazil, [28]California State University, Northridge, CA, USA, [29]University of Virginia, Charlottesville, VA, USA, [30]University of Wisconsin-Madison, Madison, WI, USA, [31]Tilburg University, Tilburg, The Netherlands, [32]Ithaca College, Ithaca, NY, USA, [33]Charles University, Prague, Czech Republic, [34]Western Kentucky University, Bowling Green, KY, USA, [35]Washington and Lee University, Lexington, VA, USA, [36]Center for Open Science, Charlottesville, VA, USA, [37]PRIME Research, Ann Arbor, MI, USA, [38]University Nijmegen, The Netherlands, [39]University of Michigan, Ann Arbor, MI, USA

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

We thank the commentators for their productive discussion of the Many Labs project (Klein et al., 2014). We entirely agree with the main theme across the commentaries: direct replication does not guarantee that the same effect was tested. As noted by Nosek and Lakens (2014, p. 137), "direct replication is the attempt to duplicate the conditions and procedure that existing theory and evidence anticipate as necessary for obtaining the effect." Attempting to do so does not guarantee success, but it does provide substantial opportunity for theoretical development building on empirical evidence.

Every replication is different in innumerable ways from the original. Evaluating high-powered replication designs a priori provides an opportunity to examine whether the theory anticipates that any of these differences will matter. Then, the experimental result informs on the theory by either (a) supporting the theory's generalizability across these presumed, and now demonstrated, irrelevant conditions, or (b) challenging the present theoretical understanding by showing that the effect does not occur under presumed irrelevant conditions, or that it does occur under conditions thought to be not amenable to obtaining the result. Finally, exploratory analysis and post facto evaluation of the outcomes provides fodder for the next iteration of theoretical development and empirical evaluation. Direct replication enables iterative cycling to refine theory and subject it to empirical confrontation.

The commentators raise relevant points on this theme in a variety of ways. Both Schwarz and Strack (2014) and Ferguson, Carter, and Hassin (2014) note the important role of theoretical analysis in the development and evaluation of a direct replication. Monin and Oppenheimer (2014) point

© 2014 Hogrefe Publishing

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

out how it is much too easy to overlook the role of stimulus selection in research design. With the pervasiveness of small sample research, this issue is difficult to address, but there is substantial opportunity to redress the limitation with larger sample research. Finally, Crisp, Miles, and Husnu (2014) note the value of aggregating evidence across investigations in order to produce the most accurate understanding of the size of an effect, rather than depending on any single demonstration.

Many Labs was a large scale replication project with many samples and settings. Nonetheless, it tested just a single operationalization of these research paradigms. It provides some definitiveness on sample and setting variation with those operationalizations, but is mute to alternative operationalizations and contexts. These commentators point out how much work is really necessary to triangulate in understanding any particular effect. Such triangulation requires more incrementalism to evaluate the boundaries and generality of an effect than is presently tolerated in peer review. A common reviewer insult is to regard a paper as incremental by "merely adding to the cumulative evidence for an effect." We hope readers will take heed of the commentators' points and appreciate the complexity of psychological effects, and the value of evaluating their reproducibility and theoretical interpretation through iterative replication designs.

Specific Reactions to Commentaries

There are some points with which we would quibble. For example: (1) Ferguson et al. suggested that other studies may have interfered with the priming, but we did not observe an effect even among those who received flag priming first (t = .339, p = .735, N = 421); and, (2) Crisp et al. suggested that a sizable portion of our sample may have been imagining an ingroup instead of an outgroup member because we did not check whether participants were Muslim – however, the portion of Muslims in the populations providing most of our samples is extremely low. Nonetheless, we were agreeable with the major themes in the commentaries, and we encourage others to explore the Many Labs dataset to inspire new hypotheses and areas for investigation (Data and materials available at: https://osf.io/wx7ck/).

Ferguson and colleagues (2014) pointed out that the predictors in the moderation model for flag priming should have been centered or standardized. We agree and thank Ferguson et al. for the correction. Table S2 (https://osf.io/v89rn/) provides the results of the hierarchical regression models estimated on standardized predictors, when all lower-order interactions and main effects are entered before the critical 3-way interaction. The two 3-way interactions testing the moderation patterns hypothesized are not different from zero.[1]

There was one point to which we respond in more detail. Schwarz and Strack (2014) suggested that the direct replications in Many Labs were only technically equivalent with no attempt in design or peer review to ensure that they were psychologically equivalent – that is, likely to engage the same psychological processes. They focused their attention on the replication of Schwarz et al. (1985), which was not altered from the original. However, we note that original materials were altered for other effects when we or reviewers deemed it important for engaging the same psychological process. For example, the original materials for the quote attribution study (Lorge & Curtiss, 1936) examined evaluations of quotes attributed to Thomas Jefferson and Vladimir Lenin, the latter target being less relevant in 2013. We changed to a new quote attributed to George Washington or Osama Bin Laden to maximize psychological equivalence. Also, we adapted the materials for the norm of reciprocity study (Hyman & Sheatsley, 1950) to refer to North Korea rather than "a Communist country like Russia."

Schwarz and Strack (2014) suggested that to conduct a direct replication of Schwarz et al. (1985) we should have altered the scale options because the original was designed presuming average television consumption of somewhat over 2 hr a day for Germans in 1983, and that Americans in 2013 watch an average of more than 5 hr per day. We did not make this change, running the risk articulated by Schwarz and Strack that the replication could be "'technically direct' while missing the goal of realizing psychological conditions that are comparable to the original study" (p. 7). However, Many Labs was not conducted on a representative sample of US adults; most samples were primarily college students.[2] Eighteen to twenty-four year olds watched approximately 3 hr of television per day in 2013 (MarketingCharts Staff, 2013), and we surmised that college students in that age range watch even less. The original scale anchors may actually be quite appropriate for this population. Further, the observed replication effect size of d = .51 almost precisely reproduced the original effect size (d = .50) leaving little evidential basis for a failure to reproduce the psychological conditions.

Schwarz and Strack (2014) also suggested that "the observed variation in effect sizes may reflect the variables that motivated the 'Many Labs' project and/or differential discrepancies between the 1983 German scale values and the actual behavioral frequencies in the samples used,

[1] During post-publication review, a discrepancy was also noticed between our replication of the Jacowitz and Kahneman (1995) anchoring procedure and the original. Ours converted a two-step item into a single response. To evaluate whether this could account for the apparently larger effect size than the original investigation, we randomly assigned Project Implicit participants to our version or the original version of the scenarios from our replication. The results indicate our version did lead to a greater effect size than the original, so this discrepancy in implementation may explain why we found a stronger anchoring effect. Full analyses and materials are available on the OSF page (see https://osf.io/wx7ck/).

[2] There are other samples in the dataset, such as highly heterogeneous MTurk and Project Implicit samples, as well as international samples that could be used to examine this issue in depth.

 © 2014 Hogrefe Publishing

which the authors decided to ignore'' (p. 7). While Schwarz and Strack are correct in principle, the variability in observed effect sizes was homogeneous ($Q_{(35)} = 36.02$, $p = .42$, $I^2 = .19$) suggesting that it could be accounted for by expected sampling error as a function of sample size.

In sum, Schwarz and Strack (2014) offered a theoretical interpretation of Schwarz et al. (1985) that highlights the potential for non (or weaker) effect size because of a presumed difference in match between scaling properties and average television watching, and anticipates heterogeneity of the effect size across samples that have different average television watching behavior. Neither of these occurred. There are two possible explanations for why we observed an effect that was nearly identical to the original finding. On one hand, the design may have induced psychological equivalence because the amount of television watched across the Many Labs samples was similar to the original study. On the other hand, this particular operationalization of the effect may not be contingent on precisely matching the scale to actual levels of behavior. Memory for the duration of activities and the frequency of habitual activities both tend to be reconstructed rather than retrieved directly and thus may be unusually malleable (Burt, 1992).

While we disagree with the particulars of the critique, we do agree with Schwarz and Strack's (2014) conceptual point – it is important that experimental manipulations engage the intended psychological process (whether in original or replication studies). It can be difficult to evaluate psychological equivalence because it is often not known which features of a design are theoretically relevant, which are relevant for correctly operationalizing a variable, and which are effectively neutral. Explicit statement of the conditions necessary to obtain a result and why these conditions are thought to matter provides opportunities to test these conditions. Replication ''successes'' and ''failures'' allow for refinement of the specifications which may have both practical and theoretical value.

Closing

We close with a word of thanks to the original authors of the effects examined in the Many Labs project. Our experience in gathering materials, soliciting feedback, and the discussion following observation of the results was positive and productive. Despite the status of replication as a central value in science, it is still a rarity in practice (Open Science Collaboration, 2012). As a consequence, it is not uncommon for original authors to feel threatened or attacked by replication efforts. None of the original authors for Many Labs responded this way. They were uniformly supportive and helpful. That does not mean that they always agreed with our decisions or interpretations, but professional disagreement is healthy for research progress. This experience may be another signal that many, perhaps most, scientists embrace the scientific norm of disinterestedness in which getting it right takes priority over one's prior claims or beliefs.

Acknowledgments

This project was supported by grants to the second and fifty-first authors from Project Implicit and by grant PRIN 2012-LATR9 N awarded to the third author. Ratliff and Nosek are consultants of Project Implicit, Inc., a nonprofit organization that includes in its mission ''to develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender, or other factors.'' RAK, MV, JC, SB, and BAN wrote the manuscript; all authors commented, edited, or approved the manuscript.

References

Burt, C. D. (1992). Reconstruction of the duration of autobiographical events. *Memory & Cognition, 20*, 124–132.

Crisp, R. J., Miles, E., & Husnu, S. (2014). Support for the replicability of imagined contact effects. Commentaries and Rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000202.

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased republican attitudes. Commentaries and Rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000202.

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In *The teaching of contemporary affairs* (pp. 11–34). New York, NY: National Council of Social Studies.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161–1166.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A ''many labs'' replication project. *Social Psychology, 45*, 142–152. doi: 0.1027/1864-9335/a000178

Lorge, I., & Curtiss, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology, 7*, 386–402.

MarketingCharts Staff. (2013). *Are young people watching less TV? (Updated - Q3 2013 Data)*. Watershed Publishing. Retrieved from http://www.marketingcharts.com/wp/television/are-young-people-watching-less-tv-24817/

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. Commentaries and Rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000202.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141. doi: 10.1027/1864-9335/a000178

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. doi: 10.1177/1745691612462588

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

© 2014 Hogrefe Publishing

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000202.

Published online May 30, 2014

Richard A. Klein

University of Florida
Department of Psychology
Gainesville, FL 32611
USA
E-mail raklein@ufl.edu

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

A New Etiquette for Replication

Daniel Kahneman

Princeton University, Princeton, NJ, USA

It is good form to pretend that science is a purely rational activity, an objective and unemotional search for the truth. But of course we all know that this image is a myth. There is a lot of passion and a lot of ego in scientists' lives, reputations matter, and feelings are easily bruised. Some interactions among scientists are fraught, and the relation between the original *author* of a piece of research and a would-be *replicator* can be particularly threatening. The purpose of this note is to propose rules for the interaction of replicators and authors, which should eventually be enforced by reviewers of proposals and reports of replication research.

I share the common position that replications play an important role in our science – to some extent by cleaning up the scientific record, mostly by deterring sloppy research. However, I believe that current norms allow replicators too much freedom to define their study as a direct replication of previous research. Authors should be guaranteed a significant role in replications of their work.

Not all replications are hostile, and many are quite friendly. However, tension is inevitable when the replicator does not believe the original findings and intends to show that a reported effect does not exist. The relationship between replicator and author is then, at best, politely adversarial. The relationship is also radically asymmetric: the replicator is in the offense, the author plays defense. The threat is one-sided because of the strong presumption in scientific discourse that more recent news is more believable. Even rumors of a failed replication cause immediate reputational damage by raising a suspicion of negligence (if not worse). The hypothesis that the failure is due to a flawed replication comes less readily to mind – except for authors and their supporters, who often feel wronged.

The difficult relationship of adversarial replication could benefit from explicit norms of conduct for both participants. One facet of the problem has already been addressed. Norms are in place to guide authors of research when they are informed that someone intends to replicate their work. They are obligated to share the details of their procedures and the entire data of their study, and to do so promptly. Unfortunately, the norms for replicators are less definite. In particular, there appear to be no rules to compel replicators to communicate with authors. Many authors have been surprised to receive, "as a courtesy," a copy of a manuscript, submitted or in press, reporting a failure to replicate one of their findings. I believe this behavior should be prohibited, not only because it is uncollegial but because it is bad science. A good-faith effort to consult with the original author should be viewed as essential to a valid replication.

In the myth of perfect science, the method section of a research report always includes enough detail to permit a direct replication. Unfortunately, this seemingly reasonable demand is rarely satisfied in psychology, because behavior is easily affected by seemingly irrelevant factors. For example, experimental instructions are commonly paraphrased in the methods section, although their wording and even the font in which they are printed are known to be significant.

It is immediately obvious that a would-be replicator must learn the details of what the author did. It is less obvious, but in my view no less important, that the original author should have detailed advance knowledge of what the replicator plans to do. The hypothesis that guides this

 © 2014 Hogrefe Publishing

proposal is that authors will generally be more sensitive than replicators to the possible effects of small discrepancies of procedure. Rules for replication should therefore ensure a serious effort to involve the author in planning the replicator's research. Of course, the rules should also be designed to prevent authors from sabotaging the replication project, as many will be tempted to do.

Here is how this proposal would work.

(1) When the replication is ready – after a pilot but before data collection – the replicator sends the author a detailed description of the planned procedure, including actual programs and a video when relevant.
(2) The author then has a limited period – perhaps a month – to respond with comments and suggested modifications of the plan.
(3) The replicator is not obliged to accept the author's suggestions, but is required to provide a full description of the final plan. The reasons for rejecting any of the author's suggestions must be explained in detail.
(4) The entire correspondence is on the record, available for subsequent reviewers to evaluate the reasonableness of the positions taken by the two sides.

The rules are designed to motivate both author and replicator to behave reasonably even when they are thoroughly irritated with each other. They know that reviewers will use the record of their interaction to assess the validity of the replication, both at the proposal stage and in the evaluation of submitted articles. They should also know that objective reviewers will not be friendly to an author who fails to respond promptly and constructively to a replication plan, or to a replicator who ignores reasonable suggestions.

Authors, whose work and reputation are at stake, should have the right to participate as advisors in the replication of their research. The obligation to consult a possibly reluctant author undoubtedly complicates life for replicators, but the burden is not crippling. Firm standards that support the active involvement of authors will contribute both to the fairness of the process and to the scientific quality of replication research.

Published online May 30, 2014

Daniel Kahneman

Woodrow Wilson School
Princeton University
322 Wallace Hall
Princeton, NJ 08544-1013
USA
E-mail kahneman@princeton.edu

https://econtent.hogrefe.com/doi/pdf/10.1027/1864-9335/a000202 - Friday, March 29, 2024 7:43:18 AM - IP Address:44.199.241.53

© 2014 Hogrefe Publishing