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

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

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.

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.

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

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

