Evaluating psychological research requires more than attention to the N: A comment on Simonsohn's (2015) "small telescopes"

Article *in* Psychological Science · May 2016

2 authors, including:

Norbert Schwarz
University of Southern California
375 PUBLICATIONS **28,785** CITATIONS

Evaluating psychological research requires more than attention to the N: A comment on Simonsohn's (2015) "small telescopes"

Norbert Schwarz [1] and Gerald L. Clore [2]
[1]University of Southern California, [2]University of Virginia

Word count: 1,494 (text without references)

Psychological Science, **in press**

We thank Klaus Fiedler, Daphna Oyserman, and Fritz Strack for discussions. Corresponding author: Norbert Schwarz, Department of Psychology, University of Southern California, 86 20 McClintock Avenue, Los Angeles, CA 90089-1061, USA; email: norbert.schwarz@usc.edu

Abstract. *Simonsohn (2015) proposed to use effect sizes of high powered replications to evaluate whether lower powered original studies could have obtained the reported effect. His focus on sample size misses that effect size comparisons are informative with regard to a theoretical question only when the replications (i) successfully realize the theoretical variable of interest, which (ii) usually requires supporting evidence from a manipulation check that should (iii) also indicate that the manipulations were of comparable strength. Because psychological phenomena are context sensitive (iv) the context of data collection should be similar and (v) the measurement procedures comparable across studies. (vi) Larger samples are often more diverse in terms of demographics and individual differences, which can further affect effect size estimates. Without attention to these points, high powered replications do not allow inferences about whether lower powered original studies could observe what they reported.*

Replications are often considered more valid than the original study when they have a larger N. Going beyond this assumption, Simonsohn (2015) proposed to use effect size estimates from high powered replications to determine whether lower powered original studies could have found what they reported: is the phenomenon seen with the “big telescope” of a large-N replication large enough to have been visible with the “small telescope” of the lower-N original study? His conceptual and methodological errors illustrate the pitfalls of a purely statistical focus.

Concepts and Manipulations

Psychologists conduct experiments to test theories. Just as original studies, replications need to ensure that the theoretically specified variables are realized. Testing feelings-as-information theory, Schwarz and Clore (1983, Experiment 2) used the first sunny days of spring after a long Midwestern winter and the inevitable return of cold, rainy weather as naturalistic mood manipulations. A mood measure confirmed more positive moods during the former than latter days. As predicted, participants evaluated their lives-as-a-whole more favorably when in a good rather than bad mood. This difference was eliminated when their attention was drawn to the weather, leading them to realize that their current feelings may not be indicative of their general quality of life. A laboratory experiment with different manipulations replicated the interaction of mood x attribution on judgments of life-satisfaction. Other work extended the theoretical rationale to the informational value of other subjective experiences, including arousal, emotions, bodily sensations, and the fluency of mental procedures (for reviews, see Schwarz & Clore, 2003, 2007).

Consistent with experimental conventions of the 1980s, Schwarz and Clore's experiment had a small N and was statistically underpowered. Simonsohn's figure 2 compares its effect size with effect sizes from two large panel surveys that assessed the covariation of sunshine on the day of interview and respondents' reports of life-satisfaction. Feddersen, Metcalfe, and Wooden (2012) found a small influence of the weather in Australia, whereas Lucas and Lawless (2013) found none in the United States. Neither of these data sets contained a mood measure, which renders them silent on whether mood influences judgments of life-satisfaction.

Simonsohn's (2015) decision to equate a conceptual variable (mood) with its manipulation (weather) is compatible with the logic of clinical trials, but not with the logic of theory testing. In clinical trials, which have inspired much of the replicability debate and its statistical focus, the operationalization (e.g., 10 mg of a drug) is itself the variable of interest; in theory testing, any given operationalization is merely one, usually imperfect, way to realize the conceptual variable. For this reason, theory tests are more compelling when the results of different operationalizations converge (Stroebe & Strack, 2014), thus ensuring that it is not "the weather" but indeed participants' (sometimes weather-induced) mood that drives the observed effect. Informative theory tests therefore require evidence that the manipulation realized the conceptual variable. Such evidence is provided by measures that assess the conceptual variable, serving as manipulation checks. Put simply, if you don't know what the mood was, you can't make inferences about the influence of mood.

Comparability and Strength of Manipulations

The size of experimental effects depends, in part, on the strength of the manipulation. Even if a manipulation successfully induced a positive mood, its observed impact will vary with the intensity of the mood. Schwarz and Clore took advantage of the upbeat affect associated with the arrival of spring in the Midwestern United States and the dread associated with a temporary return of winter. In Simonsohn's comparisons, this turns into variations in sunshine and cloud cover per se, independent of season and location. But a sunny summer day in Texas is not the psychological equivalent of a sunny spring day in the Midwest, which renders the data silent on even the most atheoretical variant of the research question: Do similar (!) weather conditions reproduce the original effect?

The comparability and strength of experimental manipulations is more often assumed than assessed. Indeed, what qualifies as sufficiently "similar" is often theoretically underspecified. Many psychological theories address how one variable (e.g., mood, motivation, attitude strength) influences another one (e.g., judgment, choice) without fully specifying the determinants of the

independent variable itself. Theories of mood and judgment, for example, are silent on what gives rise to a mood in the first place. Hence, the implementation of independent variables is frequently based on a mix of earlier results and personal intuition, further highlighting the need for sensible manipulation checks and converging evidence across different manipulations. Empirically, similarity of the procedures used in the replication and the original study is a major predictor of (non)replication. In the Open Science Collaboration's (2015) reproducibility project, 11 replications used procedures that the original authors considered inappropriate prior to data collection; 10 of them failed (Open Science Center, 2016).

Because the context sensitivity of human cognition and the dynamics of social and cultural change apply to research materials as they apply to other things psychologists study, even technically identical manipulations do not guarantee an equivalent test of the psychological phenomenon when the context changes (for extended discussions, see Fabrigar & Wegener, 2015; Schwarz & Strack, 2014). What is or is not a meaningful change in context is often controversial as the recent discussion about the fidelity of replications in the reproducibility project illustrates (Gilbert, King, Pettigrew, & Wilson, 2016; Open Science Collaboration, 2016). Nevertheless, manipulation checks that may settle the issue are routinely missing in high profile replication efforts. Empirically, the context sensitivity of a phenomenon, rated by experts who are unaware of replication results, predicts its replication likelihood (Van Bavel, Mede-Siedlecki, Brady, & Reinero, 2016) – the less context sensitive the phenomenon, the more likely it is to replicate in another lab.

Comparability of Measurement Procedures

The size of an observed effect further varies with the level of noise in its measurement. Accordingly, effect size comparisons need to attend to the comparability of the measurement procedures, which often requires attention to the psychology of self-report (Schwarz, 1999). Theories of judgment assume that the impact of a given input decreases with the number of other inputs considered in forming the judgment (Bless, Schwarz, & Wänke, 2003). For example, life-satisfaction and marital satisfaction correlate r = .32 when the questions are asked in the life-marriage order, but r = .67 when asked in the marriage-life order, reflecting that a given input has more impact when it has just been brought to mind. This impact decreases when additional relevant inputs are rendered accessible, e.g., from r = .67 to r = .46 when work satisfaction and leisure satisfaction precede the marriage and life questions (Schwarz, Strack, & Mai, 1991). Thus, identical manipulations result in smaller effects when the item of interest is preceded by other items that broaden the range of accessible inputs relevant to the judgment.

In the surveys on which Simonsohn draws, life-satisfaction was preceded by numerous other questions in interviews exceeding 80 minutes, bringing many other applicable inputs to mind. Moreover, the surveys used demographically diverse samples and spanned multiple years. In contrast, life-satisfaction and happiness where the first questions in Schwarz and Clore's experiment, conducted with a homogenous student sample on the same campus during 4 spring days in 1981. Such methodological variables affect variation in the data set and hence the observed effect size of any manipulation. They nevertheless receive little attention in prominent replication projects, which include many experiments in a single data collection. For example, the replication projects of the Open Science Center included 13 experiments in one 20-minute session for "Many Labs 1" (Klein et al., 2014), up to 15 experiments in 30 minutes for "Many Labs 2" (Klein et al., 2015), and 10 experiments in 30 minutes for "Many Labs 3" (Ebersole et al., 2015). Few, if any, of the original studies were conducted in such a format.

Conclusions

Effect size comparisons are informative with regard to a theoretical question only when the studies (i) successfully realize the theoretical variable of interest, which (ii) usually requires supporting evidence from a manipulation check that should also (iii) indicate that the manipulations were of comparable strength. Because psychological phenomena are context sensitive, merely repeating the technical moves of the original study does not guarantee the realization of comparable psychological conditions. Moreover, the manipulation of interest is not the only variable that influences the observed effect size on a given measure. Hence, (iv) the context of data collection should be similar and (v) the measurement procedures comparable across studies. (vi) Larger samples are often more diverse in terms of demographics and individual differences, which can further affect effect size estimates. Whether a replication meets the conditions of an informative comparison is best assessed by researchers with expertise in the substantive domain of study, not solely by "replication experts". Without attending to these criteria, replicators who follow Simonsohn's (2015) advice may train a big telescope with a dirty lens on the wrong planet and conclude that the original researchers' small telescope could not have discovered the (different) planet they reported on.

References

Bless, H., Schwarz, N., & Wänke, M. (2003). The size of context effects in social judgment. In J. P. Forgas, K. D. Williams, & W. von Hippel (eds.), *Social judgments: Implicit and explicit processes* (pp. 180-197). Cambridge, UK: Cambridge University Press.

Ebersole, C. R. et al. (2015). *Many Labs 3: Evaluating participant pool quality across the academic semester via replication*. Open Science Center. Retrieved from https://osf.io/csygd/

Fabrigar, L. R., & Wegener, D.T. (2015). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, http://dx.doi.org/10.1016/j.jesp.2015.07.009

Feddersen, J., Metcalfe, R., & Wooden, M. (2012). *Subjective well-being: Weather matters; climate doesn't* (Melbourne Institute Working Paper Series, 25/2012). Melbourne, Victoria, Australia: University of Melbourne. Retrieved from http://web.archive.org/web/20150107020727/http://melbourneinstitute.com/downloads/working_paper_series/wp2012n25.pdf.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*(6277), 1037-1037.

Klein, R. A., et al. (2014). Investigating variation in replicability: A ''Many Labs'' replication project. *Social Psychology*, *45*, 142–152. doi: 10.1027/1864-9335/a000178

Klein, R.A., et al. (2015). *Many Labs 2: Investigating variation in replicability across sample and setting.* Working paper. Retrieved from http://projectimplicit.net/nosek/ML2protocol.pdf

Lucas, R. E., & Lawless, N. M. (2013). Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments. *Journal of Personality and Social Psychology*, *104*, 872–884.

Open Science Center (2016). *Many Labs 5: Can conducting formal peer review in advance improve reproducibility?* Retrieved from https://docs.google.com/document/d/1tnPqr2JSpODQjJ8yB-lC6sCwuWoJn0kWEfgg8mZH5nY/edit?pref=2&pli=1

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(*6251*), aac4716.

Open Science Collaboration (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science, 351(6277),*1037-c

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.

Schwarz, N., & Clore, G.L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513 - 523.

Schwarz, N., & Clore, G.L. (2003). Mood as information: 20 years later. *Psychological Inquiry, 14*,

296-303.

Schwarz, N., & Clore, G. L. (2007). Feelings and phenomenal experiences. In A. Kruglanski & E. T. Higgins (eds.), *Social psychology. Handbook of basic principles* (2nd ed.; pp. 385-407). New York: Guilford.

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology, 45(4),* 305-306. DOI: 10.1027/1864-9335/a000202

Schwarz, N., Strack, F., & Mai, H.P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55*, 3-23.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559-569.

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

Van Bavel, J.J., Mede-Siedlecki, P., Brady, W.J., & Reinero, D.A. (2016). *Contextual sensitivity in scientific reproducibility.* NYU; manuscript under review.