Scientific replication in the study of social animals

Jay J. Van Bavel
New York University

William A. Cunningham
University of Toronto

Van Bavel, J. J., & Cunningham, W. A. (in press). Scientific replication in the study of social animals. In J. Aronson & E. Aronson. (Eds.) *The Social Animal. 12th Edition*. New York: Worth/Freeman.

Please direct correspondence to:
Jay J. Van Bavel
Department of Psychology
Center for Neural Science
New York University
6 Washington Pl, New York NY 10003
jay.vanbavel@nyu.edu

Scientific replication in the study of social animals.

It all started with a cryptic email we received from a colleague in the middle of our summer vacations: "Are you around a phone this summer? Let me know if you might have time for a short chat". After setting up a time to call, we were both left in suspense. What was so important that it could not be conveyed in an email?

When we finally spoke, our colleague Kateri McCrea asked if she could replicate one of our published papers (Cunningham, Van Bavel & Johnsen, 2008). But, she added apologetically, "I'm not trying to attack you or anything". This was funny. To a scientist, replication is like breathing. Science is a process clouded with doubt and successful replications are critical for building confidence in our findings. Failed replications root out false claims and separate science from faith. So why was our friend calling us on vacation to smooth over what should be a routine scientific practice?

Around the same time, dozens of similar conversations were taking place around the world as part of the largest replication project in the history of psychology, if not science. This massive "Reproducibility Project" was designed to estimate the *reproducibility* (the ability to reproduce the analyses performed by other scientists) and *replicability* (the ability to replicate the results with a new sample) psychology studies by re-running 100 studies published in prominent psychology journals nearly a decade earlier (OSC, 2015). The main goal was to measure the health of the field and find out which of our cherished findings was robust across time, place, and participants.

To an outsider, this should have marked an opportunity for celebration—psychology was leading the way on one of the most fundamental elements of science. By placing itself under the microscope, the field of psychology would be able to take the lead in uncovering the scientific practices and features that predict reproducibility and then reward those practices in our scientific journals and hiring decisions.

The issue was the similar, albeit less ambitious, initiatives in other fields had provided dismal results. There had been several unsuccessful attempts to replicate major findings in field as diverse as genetics, pharmacology, oncology, biology, and economics. In some fields, the

replication rate had been close to 10%, which led many to declare that science was having a "replication crisis". The crisis in faith was slowly creeping into psychology departments.

A few years earlier, one of the most eminent social psychologists had published a highly controversial paper in our field's most prestigious scientific journal, the *Journal of Personality and Social Psychology*. The paper contained experimental evidence for the existence of precognition—the conscious awareness of a future event that could not have been otherwise anticipated (Bem, 2011). This idea was the normally reserved for the pages of science fiction. But now an eminent psychologist from Cornell University was claiming that normal people possessed psychic abilities that allowed them to see into the future! In a series of shocking experiments, he reported evidence that precognition with erotic images and among people who were sensation seekers. This paper made a huge splash in the media and inspired responses that ranged from sheer awe to outright mockery.

The vast majority of scientists we knew were skeptical, if not outraged: How could such an absurd claim be published in a top scientific journal? Answer to this question led to some serious soul searching for the field. On the one hand, everyday experience suggested that ESP did not exist. If it did, gamblers would swiftly drive casinos into bankruptcy (as everyone who has ever stepped inside a casino knows, this couldn't be further from the truth).

On the other hand, none of the experimental methods used in the paper went against any of the current practices in social psychology. The paper was composed of a meticulous sequence of experiments, conducted by a highly-respected researcher. Thankfully, science is loaded with skeptics who are willing to spend their evenings and weekends trying to understand exactly what happened.

Given the universal skepticism of the initial pre-cognition finding, a wave of studies from other labs attempted to replicate the findings. They applied the same methods in the own labs—but repeatedly failed to replicate the original result (Galek et al, 2012). These failed replications sparked a serious discussion about the practices that might have produced this finding and researchers began to ask whether other surprising—if less outlandish—findings were also figments of our imagination. These conversations started in labs, graduate seminars, and conference hotel bars, but swiftly spread online with the growth of social media.

A landmark paper showed how research might be finding such shocking results. The authors presented scientific evidence for something completely absurd: listening to The Beatles "When I'm Sixty-Four" could make undergraduate students a year and a half younger! Like a magician revealing their secrets, the authors explained how they manipulated their analyses to produce such an absurd finding—a practice that is now known a "p-hacking" (Simmons, Nelson, & Simonsohn, 2011). Specifically, they explain how analysis strategies, like adjusting for irrelevant variables, using small sample sizes, and dropping certain conditions, could produce false findings. (To get a better notion of p-hacking, you can even try it yourself online: http://fivethirtyeight.com/features/science-isnt-broken/#part1). When the authors used these practices, they were able to make it look as though songs could change the reported age of young participants.

These practices are deeply problematic for science because they can create the illusion of robust, statistically significant effects that find their way into scientific journals and public discourse. As one psychologist remarked, "Researchers have been exposed to a literature that is about as representative of real science as porn movies are representative of real sex" (Lakens, 2016). In short, our science appeared to be littered with inflated effects that had been staged and edited to fit our desires rather than reflect reality.

In fact, one of the most popular findings in psychology over the past decade—the idea that posing your body into certain "high power" postures can change your hormones (Carney, Cuddy, & Yap, 2010)—was later revealed to be the result of p-hacking by the first author. The problem is that this finding was not only published in a top scientific journal, but that the work become the foundation for one of the most popular TED Talks of all time, reaching millions of viewers.

If you think back to the paper on precognition, there are a lot of decisions in that paper that look like p-hacking. For instance, the effects of precognition only worked for the erotic images and among people who were high in sensation seeking. It is not clear why precognition should only work in these specific cases and not for other stimuli or people. Thus, these analyses might have been presented in the paper because they were the only ones that produced significant results. While it is usually impossible to know if p-hacking has occurred in a given paper, there are often signs that the researchers could have use a variety of different analytic decisions to produce an effect. Perhaps the most obvious sign is that the analysis decisions are disconnected from theory,

where decisions to add other measures and dropping outliers is made on the whims of the author(s).

These papers lifted the veil on some bad practices in science and led to a critical analysis of the standard methods in psychology. Among other things, these papers highlighted the need for rigorous replications—the gold standard in science. Although scientists had been replicating their own work for decades, the top scientific journals in the field have long been reluctant to publish direct replications of previous research, especially if they failed to reproduce the key result. But independent replications by multiple labs are critical for establishing findings and building theory. If one lab finds evidence of precognition, others need to reproduce the same results before we can accept precognition as a legitimate finding. Well-executed replications in other labs are key to building consensus among scientists in other labs and incorporating new ideas into our theories. In the absence of independent replication, the results of the original study might be due to random chance, selectively publishing significant results, manipulating data, or a change in context.

This might lead one to the conclusion that psychology was on the right track by replicating a large swatch of the published literature. But inside the field, the Reproducibility Project sparked another fierce debate. Having one's work replicated is one of the most intense forms of scrutiny in science—especially in the age of social media—and peoples' professional reputations were on the line. A failure to replicate an important finding can discredit the science, as well as the scientist who originally conducted the research. And many prominent scientists were concerned that replication studies would be sloppy or incomplete—leading to a bunch of failed replications that were due to the weakness with the replication attempt rather than the original research.

In every scientific field, some findings are successfully replicated, and others are not. It is easy to know how to interpret a study that successfully reproduces that same findings as the original study—you have more convincing evidence for the effect. But interpreting "failed replications" is far more challenging. In addition to dealing with factors like random chance and the potential for p-hacking, the replication attempt might fail because there was some small error in the design or analysis, or perhaps the study was run in a very different culture or context than the original study.

Due to these interpretative ambiguity, many failed replication studies have led to fierce debates at conferences and online. Replicators often insinuate that the original researchers engaged in

sloppy or even shady practices to produce significant results. Some scientists have even made reference to the doping scandals that have plagued cycling and other sports, accusing prominent researchers of the equivalent of illegal drug use! Likewise, the researchers behind the original work have accused replicators of ill intentions or sheer incompetence. In some cases, they have even alleged that replicators engaged in a form or reverse p-hacking to engineer a replication failure. With careers and reputations at stake, these debates can get very nasty.

Many of these debates boil down to whether or not the replicators effectively created the same conditions as the original experience or whether some other factor might have led to different results. Researchers often scrutinize the original study and compare it to the replication study to determine whether the features of the original work were implemented or whether subtle factors, like the race of participants in an experiment or the geography of where the experiment was run, might account for different results. This is why a seemingly mundane practice like replication has become a major source of contention in psychology. And this is particularly true in the field of social psychology, where virtually every theory *assumes* that changes in the social context will influence behavior.

Fuel was added to this fire when international headlines screamed that a mere 39% of psychology studies in the Reproducibility Project successfully replicated the original results (OSC, 2015). The paper reported that many replication studies were not only unsuccessful, but that the effects were much weaker, on average, than the original studies. To many observers, the field of psychology had been tested and failed.

Within minutes of publication, countless scientists and journalists rushed to declare that the field of psychology was "in crisis." Others chimed in to defend the field, arguing that that the reproducibility project was a flawed and meaningless waste of time and money. Critics noted that many of the replication studies failed to re-create the conditions of the original research, making them effectively worthless (e.g., Gilbert, King, Pettigrew, & Wilson, 2016). This debate was heated and even led one author to dryly observe that psychologists were in crisis about whether or not they were in crisis (Palmer, 2016).

But was the field of psychology really doing worse than a coin flip in producing accurate knowledge? Have we learned nothing from a century of research in psychological science? Far from it. Upon closer inspection, there was a very strong relationship between the effect sizes of the original study and the replication study. In other words, manipulations that produced strong

effects in the original studies tended to produce strong effects in the replication studies roughly ten years later (see how the dots cluster along the diagonal in Figure 1). Likewise, weak effects predicted weak effects. This relationship was even stronger for successful replication (in pale yellow). Thus, even with different researchers using different samples and (often) different research materials a decade later, psychologists managed to produce very similar results.

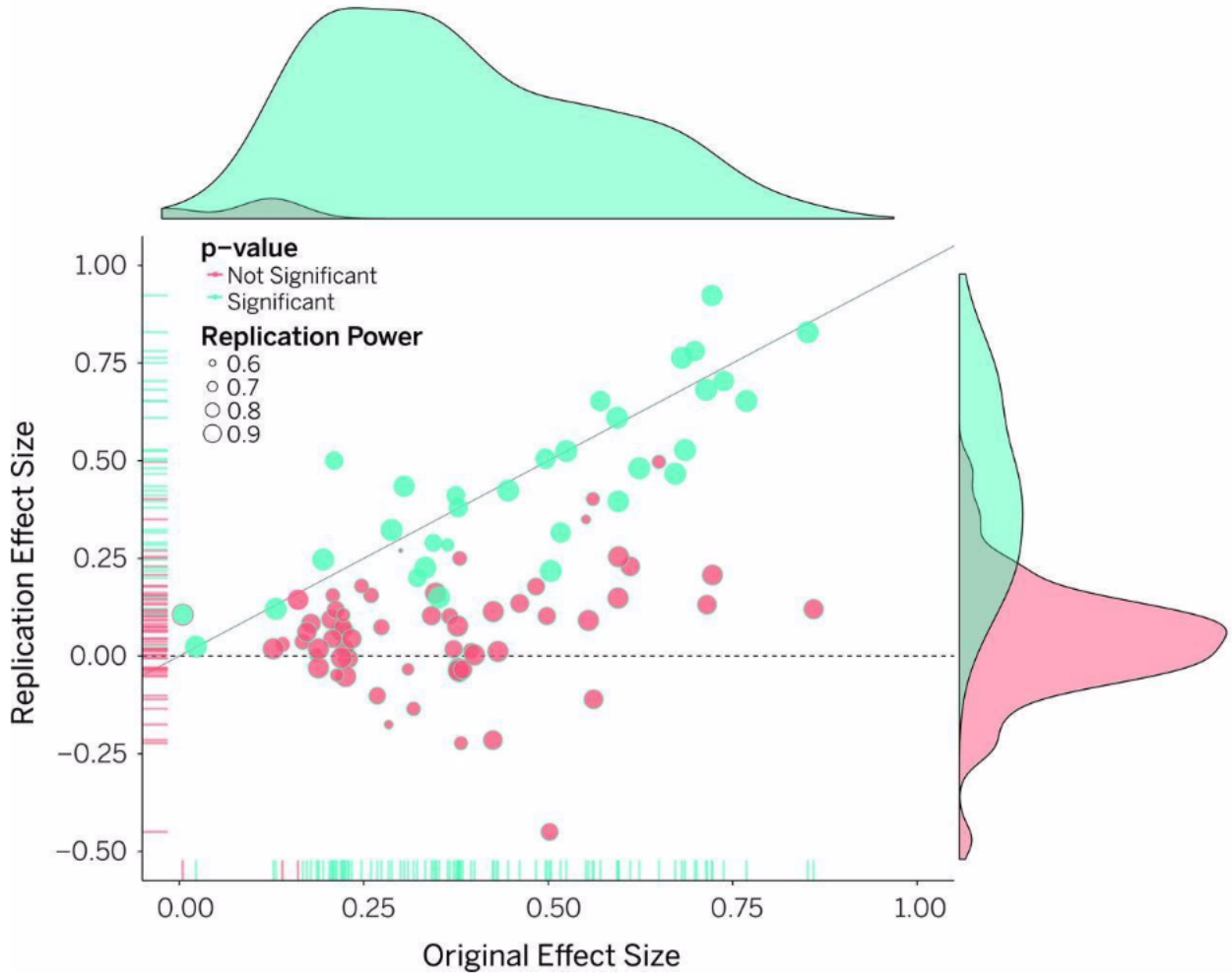

Figure 1. The relationship between the effect size of original studies (x-axis) and replication studies (y-axis). The diagonal line represents replication effect sizes that were equal to the original effect size. Each point represents a single study plus its replication. All effects below the dotted liner were in the opposite direction as the original study. The pearson correlation between the effect sizes of the original study and the replication was r = .60, suggesting that there are a strong positive relationship between the findings. However, the mean effect size of the original studies was much larger than the mean effect size of the replication studies (reprinted from Open Science Collaboration, 2015)

You might be wondering how strong was the relationship between original research and the replication results? One way of thinking about it is that the strength of the relationship between original findings and replication findings (r = .60) is very similar to personality tests: if you complete a personality test today (e.g., a measure of your extraversion or openness) you would expect to find fairly similar results ten years hence. Since a significant part of our personality is stable, we can predict our personality well into the future. Thus, even though your behavior changes from situation to situation, it still tends to be fairly predictable over time. If you were an extravert in high school there is a very good chance you will be an extravert in college. The same is true for psychology research: we can effectively predict the results of large-scale studies a decade later in the same way we can predict your personality.

What this suggests is although we all agree that replication is important, the conclusions that can be drawn following replication are not straightforward. If a study replicates, additional evidence is provided for the effect—we can infer that the effect is less likely to be a fluke than before we had the replication data. But, what are we to make of a failed replication? It often feels natural to conclude that either the original or the replication was an error. To make matter even more complicated, even if two studies are identical in stimuli and procedure, they may not be the same psychologically.

A challenging question facing the field is how to replicate studies. Replications can take multiple forms. Thus far we have been talking about *direct replications*, which aim to use the same materials and reproduce the original effect. These are fruitful for addressing the original question, but are more limited in what they can tell us about the underlying theory. This is why many psychologists favor *conceptual replications,* which aim to replicate the psychological constructs of interest, even if it requires radically departing from the specific materials used in the original study. This approach is most useful for examining whether or not the original psychological claims can generalize beyond the original materials (see Sherman & Crandall, 2016 for discussion). Both approaches to replication are crucially important to scientific progress and we cannot proceed without either of them.

The distinction between direct and conceptual replication is especially important when studying contextually sensitive research topics. It is often unclear which features of the environment might be important for producing an effect. This is why more awkward phone conversations on summer vacation might be part of the solution to the replication crisis. Our conversation with a

replicator provides a useful case study. After learning that a colleague anted to replicate our study, we immediately sent our research materials to the replication team at the University of Denver. But we also shared an important insight: they would have to completely change all the research stimuli to run the replication!

Here's why. Our original study measured emotional responses in the brain to famous people (Cunningham et al., 2008). At the time, we hypothesized that the human amygdala—a part of the temporal lobe involved in affect and emotion—would respond to motivational important stimuli (activating to positive people when we were on the lookout for positive features, and negative people when we were on the lookout for negative features). To test this idea, we presented our student participants with the names of celebrities who were expected to arouse mixed emotions (we used pilot testing to select the best list of celebrities for our study). The tricky part for the replication team was that our study was run in Canada in 2006 while the replication would be run at the University of Denver almost a decade later. If the names John A. MacDonald, Don Cherry, and Karla Homolka don't get a rise out of you, then you would have a hard time completing our study. The fact is that Canadian politicians, hockey icons, and serial killers, have little impact on the brains of most American undergraduates. Thus, it was virtually impossible to conduct a direct replication of our study a decade later and in another country.

The replication team in Denver took our advice and opted instead for a conceptual replication of the central psychological constructs (positivity, negativity and ambivalence). This required the research team to spend many additional months generating and pilot testing a new list of famous figures to use in their replication study (they were able to use half of our famous names, but had to generate half on their own). Thankfully, this painstaking process paid off—they were able to successfully replicate our findings with a much larger sample and using a number of otheanalysis strategies (Lumian & McRae, 2017). Thanks to their efforts, we now have more confidence in the conclusions from our original paper and know the findings generalize beyond Canadian students thinking about obscure Canadian celebrities!

Of course, these are not the only types of context effects that can drive differences between studies. It is impossible for a researcher to understand all the subtle manipulations that occur within their own context. One researcher may have more conservative participants on hand, whereas another may have more liberal participants. One researcher may have participants more gifted in math than another. It is possible that unmeasured factors can determine whether an

effect exists. As such, it is possible that the effect may only occur in certain situations and not in others. In other words, some additional variable may determine whether the effect is real or not. In many cases the original hypothesis is no longer valid (e.g., the effect is not universal), but there may be something interesting that can account for the discrepancy. This is important, not only for replication debates, but because we might learn something important about human nature.

To investigate whether some psychological effects may be more susceptible to changes in time or place, one of our labs coded the extent to which all of the effects reported in the original 100 studies were likely to be influenced by contextual factors such as time (e.g., pre- vs. post-Recession), culture (e.g., Eastern vs. Western culture), location (e.g., rural vs. urban setting), or population (e.g., a racially diverse population vs. a predominantly white population; Van Bavel Mende-Siedlecki, Brady, & Reinero, 2016a). We called this dimension, "context sensitivity." The coders were blind to the results of the Reproducibility Project (OSC, 2015); they had no idea when coding the studies which ones had or had not replicated.

The hypothesis was very simple: certain topics (e.g., whether cues regarding diversity signal threat or safety for African Americans) should be more sensitive to the context of the replication study than other topics, like visual statistical learning. One would assume that contextually sensitive topics, like race relations, would be *less* likely to replicate simply because it would be difficult to replicate the exact same conditions as the original study in a different time and place. After all, few people would argue that being African American is the same experience in Mississippi as it is in Montreal, or that being a woman in a computer science class in the early 1990s felt the same as it does today. Context matters to social animals, but not for all phenomena equally.

Consistent with several decades of social psychology research, the context mattered. Our ratings of contextual sensitivity predicted replication success. This was true even after statistically adjusting for methodological factors, like the sample size of the original study and replication attempt. In short, studies with higher contextual sensitivity ratings (most often the social psychological effects) were less likely to be reproduced. That said, the effects of context were modest—meaning that many other factors also predict reproducibility.

There is little question that psychologists—and other scientists—need to recruit larger and more diverse samples, share their data and materials, and find a way to publish failed replications. These factors are an important part of building a stronger science. But scientists should not ignore the fact that human behavior varies across contexts. The experience of minorities will differ dramatically between certain environments, whether we test 100 or 100,000 people. This is precisely why social psychology can provide important insights into the human condition and help better understand why some replications succeed and others fail.

Thus, it seems unlikely that the study of human behavior can ever—or should ever—aim for perfect replicability. The fact is that human behavior is incredibly complex and the study of social psychology assumes that a variety of situations will lead to different thoughts and actions. Smart scientific consumers should think critically about the conditions under which different studies were run. But the burden is also on scientists to articulate better theories and design new studies to formally test for these differences (e.g., Luttrell, Petty & Xu, 2017). This is precisely how science advances, especially important in fields like social psychology.

But these issues are far from settled. For instance, critics have noted that the reproducibility rate of social psychology (28%) is much lower than cognitive psychology (53%) (Inbar, 2016). On one hand, the authors of the Reproducibility Project even argued that the lower reproducibility rate of social psychology is due to weaker statistical power and effect sizes (OSC, 2015). On the other, fields like social psychology are interested in the power of the context—which is precisely why specific findings in that literature should vary across situations (Van Bavel, Mende-Siedlecki, Brady, Reinero, & 2016b). Indeed, large-scale international studies have found that certain findings are only replicable in the original context in which they occurred (Schweinsberg et al., 2016). Thus, while some failed replications are a sign to abandon an idea, others are an opportunity to learn more about the contextual factors that drive human behavior.

This will come as little surprise to social psychologists: The notion that human psychology is shaped by the social context has been the central premise of the field for nearly a century (Lewin, 1936). And it seems likely that this principle applies across the social sciences, from sociology to economics. There is little doubt that studying human behavior is the hardest science because we are observing the most complex of animals--ourselves. It would obviously be ideal if our greatest theorists could anticipate all the contexts in which certain relationships are likely to hold. But this noble goal is a fantasy—human behavior is far too complex. For this reason, a certain number of failed replications will be inevitable. Psychologists will need to root out flimsy effects

and faulty theories, but they should also treat them as an opportunity to learn more about human nature.

What else can psychologists do to build a better science? It would be ideal for any given finding to be explored many times by multiple labs. Surely 10 replications would give us a better sense of reality than one. Even better, it would be great to compare replications in the original context with replications conducted elsewhere to see exactly when the context matters and when it does not. Indeed, there are major initiatives in the field to do exactly this (Schweinsberg et al., 2016).

Failed replications have spawned countless important insights throughout the history of psychology. When Asian psychologists were unable to replicate many of our most cherished findings in their culture, the initial disappointment spawned influential new theories about culture. We now take for granted that there are significant differences in how individualist and collectivist cultures think, feel, and behave. Consider how unfortunate it would have been to simply dismiss a large number of American studies simply because they failed to replicate in another culture half way across the globe.

The bottom line for researchers is that all parties have a stake in working together when it comes to replication. The original researchers should share their materials, methods, and hard-earned insights to ensure the replication attempt has the best shot at success. And the replication team benefits from using and adapting these materials in a new setting. The evidence shows that replication studies that were not endorsed by the original authors were far less likely to be successful (Van Bavel et al., 2016). Thus, the best bet is usually to cooperate and learn. Even if a replication attempt fails, both parties will likely find it far more diagnostic because they agreed upon the process. Then they can set their sights on understanding why the replication results differed from the original study.

These insights are hardly limited to psychology. From Isaac Newton's prisms to contemporary research on slime molds, the history of science is full with examples of failed replications. After Newton first uncovered the light spectrum using prisms, other physicists were unable to reproduce his results. Eventually they realized the quality of glass—which was different between London and Venice—was accounting for the discrepancy. Our experiences confirm that scientists not only need better methods, but also a better understanding of context to help our replications

succeed and learn the right lessons when they fail. In either case, failed replication should inspire the development of new and improved theories about our social selves.

References

Bem, (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407-425.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21,* 1363-1368.

Cunningham, W.A., Van Bavel, J.J., & Johnsen, I.R. (2008). Affective flexibility: Evaluative processing goals shape amygdala activity. *Psychological Science, 19,* 152-160.

Fiske, S. T., & Taylor, S. E. (2013). Social cognition: From brains to culture. Sage.

Galak J., Leboeuf R. A., Nelson L. D., Simmons J. P. (2012). Correcting the past: failures to replicate ψ. *Journal of Personality and Social Psychology, 103,* 933–948

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

Inbar, Y. (2016). The association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences, 34,* E4933-4934.

Lewin, K. (1936). Principles of Topological Psychology. New York: McGraw-Hill.

Lumian, D. S. & McRae, K. (2016). Preregistered replication of "Affective flexibility: Evaluative processing goals shape amygdala activity". *Psychological Science, 28,* 1193-1200.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology, 69,* 178-183.

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

Palmer, K. M. (2016). Psychology is in crisis over whether it's in crisis. *Wired.com.* *www.wired.com/2016/03/psychology-crisis-whether-crisis/*

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., *Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C. Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., Ly, A., Marsman, M., Murase, T., Qureshi, I., Schaerer, M., Thornley, N., Tworek, C. M., Wagenmakers, E-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L.,

Kellogg, T., Kennedy, D. M., Legate, N., Luoma, T. P., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J.J. , Washburn, A. N., Wei, C., Wetter, E., Wilson, C., Darroux, S-C., & Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66, 55-67.*

Sherman & Crandall (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66,* 93-99.

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.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016a). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113,* 6454-6459.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016b). Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences, 113,* E4935-4936.