# Contextual sensitivity in scientific reproducibility

Jay J. Van Bavel[a,1], Peter Mende-Siedlecki[a], William J. Brady[a], and Diego A. Reinero[a]

[a]Department of Psychology, New York University, New York, NY 10003

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved April 25, 2016 (received for review December 3, 2015)

**In recent years, scientists have paid increasing attention to reproducibility. For example, the Reproducibility Project, a large-scale replication attempt of 100 studies published in top psychology journals found that only 39% could be unambiguously reproduced. There is a growing consensus among scientists that the lack of reproducibility in psychology and other fields stems from various methodological factors, including low statistical power, researcher's degrees of freedom, and an emphasis on publishing surprising positive results. However, there is a contentious debate about the extent to which failures to reproduce certain results might also reflect contextual differences (often termed "hidden moderators") between the original research and the replication attempt. Although psychologists have found extensive evidence that contextual factors alter behavior, some have argued that context is unlikely to influence the results of direct replications precisely because these studies use the same methods as those used in the original research. To help resolve this debate, we recoded the 100 original studies from the Reproducibility Project on the extent to which the research topic of each study was contextually sensitive. Results suggested that the contextual sensitivity of the research topic was associated with replication success, even after statistically adjusting for several methodological characteristics (e.g., statistical power, effect size). The association between contextual sensitivity and replication success did not differ across psychological subdisciplines. These results suggest that researchers, replicators, and consumers should be mindful of contextual factors that might influence a psychological process. We offer several guidelines for dealing with contextual sensitivity in reproducibility.**

replication | reproducibility | context | psychology | meta-science

In recent years, scientists have paid increasing attention to reproducibility. Unsuccessful attempts to replicate findings in genetics (1), pharmacology (2), oncology (3), biology (4), and economics (5) have given credence to previous speculation that most published research findings are false (6). Indeed, since the launch of the clinicaltrials.gov registry in 2000, which forced researchers to preregister their methods and outcome measures, the percentage of large heart-disease clinical trials reporting significant positive results plummeted from 57% to a mere 8% (7). The costs of such irreproducible preclinical research, estimated at $28 billion in the United States (8), are staggering. In a similar vein, psychologists have expressed growing concern regarding the reproducibility and validity of psychological research (e.g., refs. 9–14). This emphasis on reproducibility has produced a number of failures to replicate prominent studies, leading professional societies and government funding agencies such as the National Science Foundation to form subcommittees promoting more robust research practices (15).

The Reproducibility Project in psychology has become a landmark in the scientific reproducibility movement. To help address the issue of reproducibility in psychology, 270 researchers (Open Science Collaboration, OSC) recently attempted to directly replicate 100 studies published in top psychology journals (16). Although the effect sizes in the original studies strongly predicted the effect sizes observed in replication attempts, only 39% of psychology studies were unambiguously replicated (i.e., were subjectively rated as having replicated the original result). These findings have been interpreted as a "bleak verdict" for the state of psychological research (17). In turn, the results of the Reproducibility Project have led some to question the value of using psychology research to inform policy (e.g., ref. 18). This response corroborates recent concerns that these methodological issues in the field of psychology could weaken its credibility (19, 20).

Scientists have speculated that a lack of reproducibility in psychology, as well as in other fields, is the result of a wide range of questionable research practices, including a file-drawer problem (21, 22), low statistical power (23–25), researcher's degrees of freedom (26), presenting post hoc hypotheses as a priori hypotheses (27), and prioritizing surprising results (28, 29). In an effort to enhance the reproducibility of research, several scientific journals (e.g., *Nature* and *Science*) have offered explicit commentary and guidelines on these practices (e.g., refs. 30, 31) and have implemented new procedures, such as abolishing length restrictions on methods sections, requiring authors to affirm experimental design standards, and scrutinizing statistical analyses in consultation with statisticians. These changes are designed to increase the reproducibility of scientific results.

Many scientists have also argued that the failure to reproduce results might reflect contextual differences—often termed "hidden moderators"—between the original research and the replication attempt (32–36). In fact, such suggestions precede the current replication debate by decades. In 1981, social psychologist John Touhey criticized a failed replication of his research based on the "dubious ... assumption that experimental manipulations can be studied apart from the cultural and historical contexts that define their meanings" (p. 594 in ref. 37). Indeed, the insight that behavior is a function of both the person and the environment—elegantly captured by Lewin's equation: $B = f(P,E)$ (38)—has shaped the direction of social psychological research for more than half a century. During that time, psychologists and other social scientists have paid considerable attention to the influence of context on the individual (e.g., refs. 39–42) and have found extensive evidence that contextual factors alter human behavior (43–46).

Understanding contextual influences on behavior is not usually considered an artifact or a nuisance variable but rather can be a driving force behind scientific inquiry and discovery. As statistician and political scientist Andrew Gelman recently

## Significance

**Scientific progress requires that findings can be reproduced by other scientists. However, there is widespread debate in psychology (and other fields) about how to interpret failed replications. Many have argued that contextual factors might account for several of these failed replications. We analyzed 100 replication attempts in psychology and found that the extent to which the research topic was likely to be contextually sensitive (varying in time, culture, or location) was associated with replication success. This relationship remained a significant predictor of replication success even after adjusting for characteristics of the original and replication studies that previously had been associated with replication success (e.g., effect size, statistical power). We offer recommendations for psychologists and other scientists interested in reproducibility.**

Author contributions: J.J.V.B., P.M.-S., W.J.B., and D.A.R. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data are available on the Open Science Framework.

[1]To whom correspondence should be addressed. Email: jay.vanbavel@nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1521897113/-/DCSupplemental.

suggested, "Once we realize that effects are contextually bound, a next step is to study how they vary" (33). Indeed, the OSC authors correctly note that "there is no such thing as exact replication" in the field of psychology (47). Although the ideal methods section should include enough detail to permit a direct replication, this seemingly reasonable demand is rarely satisfied in psychology, because human behavior is easily affected by seemingly irrelevant factors (48).

The issue of hidden moderators is not limited to psychology. For instance, many rodent studies are doomed to irreproducibility because subtle environmental differences, such as food, bedding, and light, can affect biological and chemical processes that determine whether experimental treatments succeed or fail (49). Likewise, Sir Isaac Newton alleged that his contemporaries were unable to replicate his research on the color spectrum of light because of bad prisms (50). After he directed his contemporaries to better prisms (ones produced in London rather than in Italy), they were able to reproduce his results. Thus the contextual differences between the conditions in which initial and replication studies are conducted appear to influence reproducibility across scientific disciplines, ranging from psychology to biology to physics.

Although the notion that "context matters" is informally acknowledged by most scientists, making this common sense assumption explicit is important because the issue is fundamental to most research (51). Indeed, the role of context is frequently overlooked—and even dismissed—in the evaluation of replication results. Several scientists have argued that hidden moderators such as context are unlikely to influence the results of direct replications, precisely because the replication studies use the same methods used in the original research (52, 53). Similarly, others have argued that direct replications are the strongest (and possibly only) believable evidence for the reliability of an effect (54, 55). This approach calls into question the influence of hidden moderators.

This issue is especially contentious in psychology because replication attempts inevitably differ from the original studies. For instance, a recent critique of the Reproducibility Project alleged that several replication studies differed significantly from the original studies, undercutting any inferences about lack of reproducibility in psychology (56). The allegation that low-fidelity replication attempts undercut the validity of the Reproducibility Project launched a debate about the role of contextual factors in several replication failures, both in print (47) and in subsequent online commentaries (e.g., refs. 57–59). According to a Bayesian reanalysis of the Reproducibility Project, one pair of authors argued that "the apparent discrepancy between the original set of results and the outcome of the Reproducibility Project can be explained adequately by the combination of deleterious publication practices and weak standards of evidence, without recourse to hypothetical hidden moderators" (60). However, this paper did not directly code or analyze contextual sensitivity in any systematic way. Despite the centrality of this issue for interpreting scientific results in psychology and beyond, very little research has empirically examined the role of contextual sensitivity in reproducibility.

Among the few efforts to examine the relationship between context and reproducibility in psychology, the results have been mixed. One large-scale replication tested 13 effects (10 were reproduced consistently, and one was reproduced weakly) across 36 international samples (61). They observed only small effects of setting and a much stronger influence of the effects themselves (i.e., some effects are simply more robust than others, regardless of setting).[†] The authors concluded that context (i.e., sample/setting) had "little systematic effect on the observed results" (51). In contrast, a project examining the reproducibility of 10 effects related to moral judgment (seven were reproduced consistently and one was reproduced weakly) across 25 international samples (62) found evidence that certain effects were reproducible only within the culture in which they were originally observed. In other words, context moderated replication success.

The relatively small number of replication attempts (along with relatively idiosyncratic inclusion criteria[‡]) across these prior replication projects makes it difficult to draw strong conclusions regarding the role of contextual sensitivity in reproducibility. Furthermore, if the effects chosen for replication in these projects were predominantly effects which are a priori unlikely to vary by context, then it would come as no surprise that context does not predict replication success. This paper addresses these issues directly by analyzing a large and diverse database of 100 replication attempts and assessing the contextual sensitivity of each effect.

## Methods

To help assess the relationship between context and replication success, we coded and analyzed contextual sensitivity in the Reproducibility Project (16). Three coders with graduate training in psychology (one postdoctoral coder and two predoctoral students with experience in social, cognitive, and neuroscience laboratories; their professional credentials are publicly available at https://osf.io/cgur9/) rated the 100 original studies presented in the Reproducibility Project (16) on the extent to which the research topic in each study was contextually sensitive. The raters were unaware of the results of replication attempts. Before coding any studies, the coders practiced their rating scheme on an independent set of four studies addressed in other replication efforts (63–66). This practice ensured that each coder rated contextual sensitivity in a similar and consistent fashion. Once consistency was established, the three coders moved on to the 100 studies contained in the Reproducibility Project.

Twenty-five of these studies were randomly selected to be rated by all three coders so that a measure of interrater reliability could be computed. Each coder also rated a distinct set of 25 randomly assigned studies independently, bringing each coder's total number of rated studies to 50. When rating a study, the coder assessed how likely the effect reported in the abstract of the original study was to vary by context—defined broadly as differing in time (e.g., pre- vs. post-Recession), culture (e.g., individualistic vs. collectivistic culture), location (e.g., rural vs. urban setting), or population (e.g., a racially diverse population vs. a predominantly White population). This coding scheme concerned broad classes of macrolevel contextual influences that could reasonably be expected to influence the reproducibility of psychological research.

Coders did not attempt to make explicit predictions about whether the specific replication attempt in question would succeed, nor did they attempt to make judgments about the quality of the original research. Moreover, coders did not base their assessments of contextual sensitivity on the reputations of particular laboratories, researchers, or effects, nor did they assess objective information regarding subsequent replication attempts available in the literature. Rather, coders were tasked solely with evaluating the likelihood that a given effect might fluctuate if a direct replication was conducted outside the original context in which it was obtained. In the few cases in which the original articles did not contain abstracts (5 of 100 studies), coders inspected the methods section of that study in the original article. In addition to the coders being largely blind to methodological factors associated with reproducibility, we statistically adjusted for several of these factors in regression models reported below.

Contextual sensitivity ratings were made on a five-point scale, with anchors at 1 (context is not at all likely to affect results), 3 (context is somewhat likely to affect results), and 5 (context is very likely to affect results) (mean = 2.90, SD = 1.16). Reliability across raters was high: An intraclass correlation test for consistency revealed an alpha of 0.86 for the subset of 25 studies reviewed by all three coders [intraclass correlation coefficients (2,3)] (67). For instance, context was expected to be largely irrelevant for research on visual statistical learning (rated 1) (68) or for the action-based model of cognitive dissonance (rated 2) (69), was expected to have some influence on research concerning bilingualism and inhibitory control (rated 3) (70), and was expected to have a significant impact on research on the ultimate sampling dilemma (rated 4) (71) and on whether cues regarding diversity signal threat or safety for African Americans (rated 5) (72).

[†]It is worth noting that Many Labs 1 (61) found considerable heterogeneity of effect sizes for nearly half of their effects (6/13). Furthermore, they found sample (United States vs. international) and setting (online vs. in-lab) differences for nearly one-third (10/32) of their moderation tests, seven of which were among the largest effects (i.e., anchoring, allowed–forbidden). As the authors note, one might expect such contextual differences to arise for anchoring effects because of differences between the samples in knowledge such as the height of Mt. Everest, the distance to New York City, or the population of Chicago. Thus, context did indeed have a systematic effect on the observed results.

[‡]Twelve of the 13 studies presented by Many Labs 1 were selected for the project based on criteria that included suitability for online presentation (e.g., to allow comparisons between online and in-lab samples), study length, and study design (i.e., only simple, two-condition designs were included, with the exception of one correlational study).

Satisfied that contextual sensitivity ratings were consistent across raters, we computed a simple average of those ratings for the subset of studies reviewed by all three coders and assessed the degree to which contextual sensitivity covaried with replication success (*Additional Analysis Details of Coder Variability*). We then compared the effect of contextual sensitivity on replication success relative to other variables that have been invoked to explain replication success or failure (materials, data, and further analysis details are available online at https://osf.io/cgur9/). This procedure was a conservative test of the role of contextual sensitivity in reproducibility because most replications were explicitly designed to be as similar to the original research as possible. In many cases (80/100), the original authors evaluated the appropriateness of the methods before data collection. We also explicitly compared these replication attempts with those in which the authors explicitly preregistered concerns about the methods before data collection.

All regression analyses reported below were conducted using either a binary logistic regression model or linear regression models (see *Multiple Regression Parameters*). We analyzed two models. In model 1, we included the contextual sensitivity variable as well as four other variables that were found to predict subjective replication success by the OSC (16): (*i*) the effect size of the original study; (*ii*) whether the original result was surprising, as coded by Reproducibility Project coordinators on a six-point scale in response to the question "To what extent is the key effect a surprising or counterintuitive outcome?" ranging from 1 (not at all surprising) to 6 (extremely surprising); (*iii*) the power of the replication attempt; and (*iv*) whether the replication result was surprising, coded by the replication team on a five-point scale in response to the question "To what extent was the replication team surprised by the replication results?" ranging from 1 (results were exactly as anticipated) to 5 (results were extremely surprising) (Table S1). In model 2 we included these five variables and two other variables that are widely believed to influence reproducibility: (*i*) the sample size of the original study and (*ii*) the similarity of the replication as self-assessed by replication teams on a seven-point scale in response to the question "Overall, how much did the replication methodology resemble the original study?" ranging from 1 (not at all similar) to 7 (essentially identical) (Table S2 and see Table S3 for full correlation matrix of contextual variability with original and replication study characteristics).

## Results

The results confirmed that contextual sensitivity is associated with reproducibility. Specifically, contextual sensitivity was negatively correlated with the success of the replication attempt, $r(98) = -0.23$, $P = 0.024$ (Table S4), such that the more contextually sensitive a topic was rated, the less likely was the replication attempt to be successful.[§] We focused on the subjective binary rating of replication success as our key dependent variable of interest because it was widely cited as the central index of reproducibility, including in immediate news reports in *Science* (74) and *Nature* (75). Nevertheless, we reanalyzed the results with all measures of reproducibility [e.g., confidence intervals (CI) and meta-analysis] and found that the average correlation between contextual sensitivity and reproducibility was virtually identical to the estimate we found with the subjective binary rating (mean $r = -0.22$). As such, the effect size estimate appeared to be relatively robust across reproducibility indices (Table S4).

We then compared the effects of contextual sensitivity with other research practices that have been invoked to explain reproducibility. Multiple logistic regression analysis conducted for model 1 indicated that contextual sensitivity remained a significant predictor of replication success, $B = -0.80$, $P = 0.015$, even after adjusting for characteristics of the original and replication studies that previously were associated with replication success (5), including: (*i*) the effect size of the original study, (*ii*) whether the original result was surprising, (*iii*) the power of the replication attempt, and (*iv*) whether the replication result was surprising. Further, when these variables were entered in the first step of a hierarchical regression, and contextual sensitivity was entered in the second step, the model with contextual sensitivity was a significantly better fit for the data, $\Delta R^2 = 0.06$, $P = 0.008$ (Table S1). Thus, contextual sensitivity provides incremental predictive information about reproducibility.

Although the Reproducibility Project did not observe that original sample size was a significant predictor of replication success (5), several studies have suggested that sample size may constitute a critical methodological influence on reproducibility (23–25). In addition, a common critique of recent replication attempts is that many such studies do not actually constitute direct replications, and there may be a number of more obvious moderators (e.g., using different materials) (56, 76). To assess the role of these potential moderators, we entered these additional variables in a second regression model (model 2) and observed that the effect of contextual sensitivity continued to be significantly associated with replication success ($P = 0.023$), even when adjusting for the sample size of the original study and the similarity of the replication, in addition to the study characteristics entered in model 1 (Fig. 1 and Table S2). This result suggests that contextual sensitivity plays a key role in replication success, over and above several other important methodological characteristics.

We also examined whether the relationship between contextual sensitivity and reproducibility was specific to social psychology studies.[¶] Although social psychology studies (mean = 3.58) were rated as much more contextually sensitive than cognitive psychology studies (mean = 2.00), $t(98) = -9.14$, $P < 0.001$, $d = -1.85$, we found no evidence of an interaction between contextual sensitivity and subdiscipline on replication success ($P = 0.877$). Furthermore, the effect size for contextual sensitivity predicting replication success was nearly identical for social psychology studies and cognitive psychology studies. (In a binary logistic regression predicting replication success, we entered contextual sensitivity, subdiscipline, and their interaction as predictors. Simple effects analysis via dummy coding demonstrated that the effect of contextual sensitivity was nearly identical in magnitude for social psychology studies [odds ratio (OR) = 0.823] and cognitive psychology studies [OR = 0.892]) In other words, contextual sensitivity appears to play an important role in replication success across multiple areas of psychology, and there is good reason to believe the effect of contextual sensitivity applies to other scientific fields as well (80, 81).

To elaborate further on the role of contextual sensitivity in reproducibility, we report the results of the original researchers who did or did not express concerns about the design of the replication study. The OSC asked original researchers to comment on the replication plan and coded their responses as 1 = endorsement; 2 = concerns based on informed judgment/speculation; 3 = concerns based on unpublished empirical evidence of the constraints on the effect; 4 = concerns based on published empirical evidence of the constraints on the effect; and 9 = no response. We compared replications that were endorsed by the original authors ($n = 69$) with replications for which the original authors had preregistered concerns ($n = 11$). Eighteen original authors did not reply to replicators' requests for commentary, and two replication teams did not attempt to contact original authors.

Of the 11 studies in which the original authors explicitly preregistered some sort of concern, the original results could be reproduced successfully in only one (9%). This subset of 11 studies was rated higher on contextual sensitivity (mean = 3.73) than studies in which the original researchers expected their results to be replicated successfully or in which they did not make a prediction about replication (mean = 2.80, $P < 0.001$). Although there are numerous reasons an author may express concerns about

[§]A meta-analysis of 322 meta-analytic investigations of social psychological phenomena observed that the average effect in social psychology had an effect size of $r = 0.21$ (73). It is interesting that a similar effect of context holds even in this set of direct replication attempts.

[¶]To our knowledge, the categorization of studies in the Reproducibility Project followed the following process: A graduate student involved in the Reproducibility Project initially categorized all 100 studies according to subfield (e.g., social vs. cognitive psychology). Replication teams could then recode those categorizations (although it is not clear whether recoding was done or, if so, which studies were recoded). Finally, Brian Nosek reviewed the categorizations. These categorizations possess a high degree of face validity, but this scheme may have resulted in a handful of contentious assignments. For example, studies on error-related negativity [specifically associated with performance in the Eriksen flanker task (77)], the value heuristic [i.e., the propensity to judge the frequency of a class of objects based on the objects' subjective value (78)], and conceptual fluency (79) were all classified as social psychology studies rather than as cognitive psychology studies.

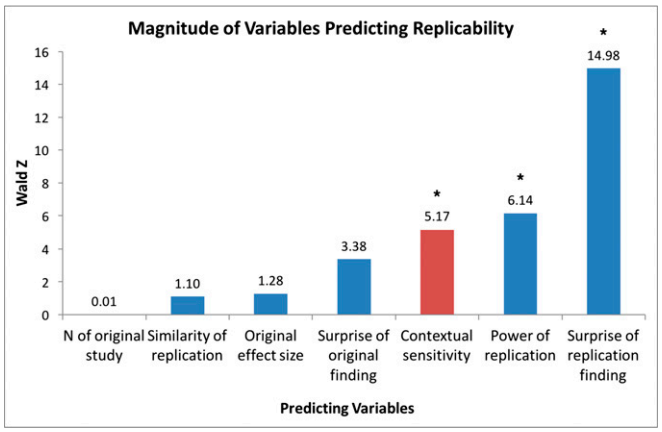


**Fig. 1.** The magnitude (Wald *Z*) of variables previously associated with replication success (blue) and contextual sensitivity (red) when entered simultaneously into a multiple logistic regression (with subjective replication success (yes or no) as the binary response variable). Contextual sensitivity, power of replication, and surprisingness of replication finding (as rated by the replication team) remained significant predictors of replication success; $*P < 0.05$.

the replication design (47), these specific 11 studies appeared to involve highly contextually sensitive topics. We suspect the authors may have understood this characteristic of their research topic when they expressed concerns. However, some people have speculated that the original authors preregistered concern only because they were aware that their studies were relatively weak based on other factors affecting replication (e.g., small effect sizes, underpowered designs) (59). The OSC authors also argued that authors who were less confident of their study's robustness may have been less likely to endorse the replications (47).

To discern between these two competing alternatives, we ran a binary logistic regression predicting whether authors would express concern about the replication attempt. Strikingly, we found that when study characteristics associated with replication were entered into the model along with the contextual sensitivity variable, the only significant predictor of whether an author expressed concern was contextual sensitivity, $B = 1.30$, $P = 0.004$. [All study characteristics used in model 2 were input into the regression as simultaneous predictors.] Thus, as the contextual sensitivity of their effect increased, authors were 3.68 times more likely to express concern. Expressing concern was not correlated with the other key study characteristics ($P$s > 0.241). The results from this relatively small sample of studies should be interpreted cautiously until more data can be collected. However, they suggest that original authors may be attuned to the potential problems with replication designs and that these concerns do not appear to derive from methodological weaknesses in the original studies.

The endorsement of the original authors also predicted replication success. Specifically, a Pearson's $\chi^2$ confirmed that the replication rate of the studies in which the original authors endorsed the replication study (46%) was more than five times higher than in the studies for which the original authors expressed concerns (9%; $\chi^2 = 4.01$, $P = 0.045$). This result suggests that author endorsement effectively predicts future replication success. Moreover, when the 11 studies about which the original authors expressed concerns were removed, the effect sizes in the remaining original studies were highly correlated with the effect sizes observed in the replication studies (Pearson's $r = 0.60$).[#] As such, there appears to be a strong correlation between the original findings and results of the replication. Taken together, these results suggest that replication success is higher when the original authors endorse the design of replication studies, and the impact of endorsement appears to be most relevant when scientists are trying to replicate contextually sensitive effects.

## Discussion

This paper provides evidence that contextual factors are associated with reproducibility, even after adjusting for other methodological variables reported or hypothesized to impact replication success. Attempting a replication in a different time or place or with a different sample can alter the results of what are otherwise considered "direct replications." The results suggest that many variables in psychology and other social sciences cannot be fully understood apart from the cultural and historical contexts that define their meanings (37).

Our findings raise a number of questions about how the field might move forward in the face of a failed replication. We submit that failed replication attempts represent an opportunity to consider new moderators, even ones that may have been obscure to the original researchers, and to test these hypotheses formally (34). According to William McGuire, "empirical confrontation is a discovery process to make clear the meaning of the hypothesis, disclosing its hidden assumptions and thus clarifying circumstances under which the hypothesis is true and those under which it is false" (34). Indeed, many scientific discoveries can be traced to a failed replication (32), and entire fields are built on the premise that certain phenomena are bound by cultural or other contextual factors.

Moreover, our results suggest that experts are able to identify factors that will influence reproducibility and that original researchers seem to be attuned to these factors when evaluating replication designs. However, it is important to note that contextual sensitivity does not necessarily suggest a lack of robustness or reproducibility. For instance, contextual variation is itself incredibly robust in some areas of research (73). Furthermore, contextual sensitivity is sufficient but not necessary for variation in the likelihood of replication. A number of other methodological characteristics in a given study may be associated with a failure to replicate (16). However even a large effect in a methodologically sound study can fail to replicate if the context is significantly different, and in many cases the direction of the original effect can even be reversed in a new context.

Given these considerations, it may be more fruitful to empirically and theoretically address failed replications than debate whether or not the field is in the midst of a "replication crisis." At the same time, hidden moderators should not be blindly invoked as explanations for failed replications without a measure of scrutiny. To forestall these concerns, we encourage authors to share their research materials, to avoid making universal generalizations from limited data, to be as explicit as possible in defining likely contextual boundaries on individual effects, and to assess those boundaries across multiple studies (40, 82, 83). Psychologists should also avoid making mechanistic claims, as this approach necessitates that manipulating one variable always and exclusively leads to a specific, deterministic change in another, precluding the possibility of contextual influence (84). Psychological variables almost never involve this form of deterministic causation, and suggesting otherwise may lead replicators and the public to infer erroneously that a given effect is mechanistic. By following these guidelines, scientists acknowledge potential moderating factors, clarify their theoretical framework, and provide a better roadmap for future research (including replications).

We advocate that replicators work closely with original researchers whenever possible, because doing so is likely to improve the rate of reproducibility (see *Additional Data Advocating for Consultation with Original Authors*), especially when the topic is likely to be contextually sensitive. As Daniel Kahneman recently suggested, "A good-faith effort to consult with the original author should be viewed as essential to a valid replication ... .

[#]The correlation between the effect sizes in the remaining original studies strongly predicted the effect sizes observed in replication attempts, and this correlation was nearly identical when we include all 100 studies (Pearson's $r = 0.60$). As such, this correlation cannot be attributed to the removal of the 11 studies about which the authors expressed concerns (although the correlation within these 11 studies is only $r = 0.18$). We report it here for completeness.

The hypothesis that guides this 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" (48). Our data appear to bear out this suggestion: Original researchers seem capable of identifying issues in the design of replication studies, especially when these topics are contextually sensitive, and the replications of studies about which the researchers have such concerns are highly unlikely to be successful. This sort of active dialogue between replicators and original researchers is also at the core of a recently published "replication recipe" attempting to establish standard criteria for a "convincingly close replication" (76).

Ultimately, original researchers and replicators should focus squarely on psychological process. In many instances, the original research materials may be poorly suited for eliciting the same psychological process in a different time or place. When a research topic appears highly sensitive to contextual factors, conceptual replications offer an important alternative to direct replications. In addition to assessing the generalizability of certain results, marked departures from the original materials may be necessary to elicit the psychological process of interest. In this way, conceptual replications can even improve the probability of successful replication (however, see ref. 85 for falsifiability limitations of conceptual replications).

We wholeheartedly agree that publication practices and methodological improvements (e.g., increasing power, publishing nonsignificant results) are necessary for improving reproducibility. Indeed, our analyses support these claims: The variance explained by contextual sensitivity is surpassed by the statistical power of the replication attempt. Numerous other suggestions for improving reproducibility have been proposed (e.g., refs. 62, 76, 86, 87). For example, the replication recipe (76) offers a "five-ingredient" approach to standardizing replication attempts that emphasizes precision, power, transparency, and collaboration. However, our findings suggest that these initiatives are no substitute for careful attention to psychological process and the context in which the original and replication research occurred.[||]

Researchers, replicators, and consumers must be mindful of contextual factors that might influence a psychological process and seek to understand the boundaries of a given effect. After all, the brain, behavior, and society are orderly in their complexity rather than lawful in their simplicity (88, 89). It is precisely because of this complexity that psychologists must grapple with contextual moderators. Although context matters across the sciences (e.g., humidity levels in a laboratory unexpectedly influencing research on the human genome), psychologists may be in a unique position to address these issues and apply these lessons to issues of reproducibility. By focusing on why some effects appear to exist under certain conditions and not others, we can advance our understanding of the boundaries of our effects as well as enrich the broader scientific discourse on reproducibility.

Our research represents one step in this direction. We found that the contextual sensitivity of research topics in psychology was associated with replication success, even after statistically adjusting for several methodological characteristics. This analysis focused on broad, macrolevel contextual influences—time, culture, location, and population—and, ultimately, collapsed across these very different sources of variability. Future work should test these (and other) factors separately and begin to develop a more nuanced model of the influence of context on reproducibility. Moreover, the breadth of contextual sensitivity surveyed in our analysis might represent an underestimation of a host of local influences that may determine whether an effect is replicated. These additional influences range from obvious but sometimes overlooked factors, such as the race or gender of an experimenter (90), temperature (91), and time of day (92), to the more amorphous (e.g., how the demeanor of an experimenter conducting a first-time test of a hypothesis she believes is credible may differ from that of an experimenter assessing whether a study will replicate). Although it is difficult for any single researcher to anticipate and specify every potential moderator, that is the central enterprise of future research. The lesson here is not that context is too hard to study but rather that context is too important to ignore.

**ACKNOWLEDGMENTS.** We thank the Reproducibility Project for graciously making these data available; and Lisa Feldman Barrett, Gerald Clore, Carsten De Drue, Mickey Inzlicht, John Jost, Chris Loersch, Brian Nosek, Dominic Packer, Bernadette Park, Rich Petty, and three anonymous reviewers for helpful comments on this paper. This work was supported by National Science Foundation Grant 1555131 (to J.J.V.B.).

[||]Our results have important implications for journal editors hoping to enact explicit replication recommendations for contributing authors. For example, in a recent editorial, *Psychological Science* editor Steven Lindsay wrote, "Editors at *Psychological Science* are on the lookout for this troubling trio: (a) low statistical power, (b) a surprising result, and (c) a *p* value only slightly less than .05. In my view, *Psychological Science* should not publish any single-experiment report with these three features because the results are of questionable replicability." (31). Although we side with the editor on this matter, context may have as much predictive utility as any individual component of this "troubling trio."

1. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* 4(2):45–61.
2. Prinz F, Schlange T, Asadullah K (2011) Believe it or not: How much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 10(9):712.
3. Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531–533.
4. Reaves ML, Sinha S, Rabinowitz JD, Kruglyak L, Redfield RJ (2012) Absence of detectable arsenate in DNA from arsenate-grown GFAJ-1 cells. *Science* 337(6093):470–473.
5. Chang AC, Li P (2015) *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"*, Finance and Economics Discussion Series 2015-083. (Board of Governors of the Federal Reserve System, Washington, DC).
6. Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8):e124.
7. Kaplan RM, Irvin VL (2015) Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One* 10(8):e0132382.
8. Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. *PLoS Biol* 13(6):e1002165.
9. Bakker M, Wicherts JM (2011) The (mis)reporting of statistical results in psychology journals. *Behav Res Methods* 43(3):666–678.
10. Fiedler K (2011) Voodoo correlations are everywhere—not only in neuroscience. *Perspect Psychol Sci* 6(2):163–171.
11. García-Pérez MA (2012) Statistical conclusion validity: Some common threats and simple remedies. *Front Psychol* 3:325.
12. John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci* 23(5):524–532.
13. Pashler H, Wagenmakers EJ (2012) Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspect Psychol Sci* 7(6): 528–530.
14. Spellman BA (2012b) Introduction to the special section on research practices. *Perspect Psychol Sci* 7(6):655–656.
15. Bollen K, Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL (2015) Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science: Report of the Subcommittee on Replicability in Science, Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. Available at www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf. Accessed May 10, 2016.
16. Open Science Collaboration (2015) Psychology. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.
17. Sample I (2015) Study delivers bleak verdict on validity of psychology experiment results. Available at https://www.theguardian.com/science/2015/aug/27/study-delivers-bleak-verdict-on-validity-of-psychology-experiment-results. Accessed May 10, 2016.
18. Efferson ADP (2015) How many laws are based on psychology's bad science. Available at thefederalist.com/2015/09/08/how-many-laws-are-based-on-psychologys-bad-science/. Accessed May 10, 2016.
19. Ferguson CJ (2015) "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *Am Psychol* 70(6):527–542.
20. Lilienfeld SO (2012) Public skepticism of psychology: Why many people perceive the study of human behavior as unscientific. *Am Psychol* 67(2):111–129.
21. Ferguson CJ, Heene M (2012) A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspect Psychol Sci* 7(6):555–561.
22. Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychol Bull* 86(3):638–641.
23. Cohen J (1962) The statistical power of abnormal-social psychological research: A review. *J Abnorm Soc Psychol* 65:145–153.
24. Maxwell SE, Lau MY, Howard GS (2015) Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am Psychol* 70(6):487–498.

25. Vankov I, Bowers J, Munafò MR (2014) On the persistence of low power in psychological science. *Q J Exp Psychol (Hove)* 67(5):1037–1040.
26. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22(11):1359–1366.
27. Kerr NL (1998) HARKing: Hypothesizing after the results are known. *Pers Soc Psychol Rev* 2(3):196–217.
28. Yong E (2012) Replication studies: Bad copy. *Nature* 485(7398):298–300.
29. Young NS, Ioannidis JPA, Al-Ubaydli O (2008) Why current publication practices may distort science. *PLoS Med* 5(10):e201.
30. Cumming G (2014) The new statistics: Why and how. *Psychol Sci* 25(1):7–29.
31. Lindsay DS (2015) Replication in psychological science. *Psychol Sci* 26(12):1827–1832.
32. Feldman-Barrett L (2015) Psychology is not in crisis. Available at www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html?_r=1. Accessed May 10, 2016.
33. Gelman A (2014) The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *J Manage* 41(2):632–643.
34. McGuire WJ (2013) An additional future for psychological science. *Perspect Psychol Sci* 8(4):414–423.
35. Klein O, et al. (2012) Low hopes, high expectations expectancy effects and the replicability of behavioral experiments. *Perspect Psychol Sci* 7(6):572–584.
36. Stroebe W, Strack F (2014) The alleged crisis and the illusion of exact replication. *Perspect Psychol Sci* 9(1):59–71.
37. Touhey JC (1981) Replication failures in personality and social psychology negative findings or mistaken assumptions? *Pers Soc Psychol Bull* 7(4):593–595.
38. Lewin K (1936) *Principles of Topological Psychology* (McGraw-Hill, New York) trans Heider F and Heider G.
39. Mischel W (1977) The interaction of person and situation. *Personality at the Crossroads: Current Issues in Interactional Psychology*, eds Magnnusson D, Endler D (Lawrence Erlbaum Associates, Hillsdale, NJ), pp 333–352.
40. Rousseau DM, Fried Y (2001) Location, location, location: Contextualizing organizational research. *J Organ Behav* 22(1):1–13.
41. Sarason IG, Smith RE, Diener E (1975) Personality research: Components of variance attributable to the person and the situation. *J Pers Soc Psychol* 32(2):199–204.
42. Weick KE (1996) Enactment and the boundaryless career. *The Boundaryless Career: A New Employment Principle for a New Organizational Era*, eds Arthur MB, Rousseau DM (Oxford Univ Press, New York), pp 40–57.
43. Camerer CF, Loewenstein G, Rabin M, eds (2011) *Advances in Behavioral Economics* (Princeton Univ Press, Princeton, NJ).
44. Fiske ST, Gilbert DT, Lindzey G (2010) *Handbook of Social Psychology* (John Wiley & Sons, Hoboken, NJ).
45. Goodin RE, ed (2009) *The Oxford Handbook of Political Science* (Oxford Univ Press, New York).
46. Hedström P, Bearman P, eds (2009) *The Oxford Handbook of Analytical Sociology* (Oxford Univ Press, New York).
47. Anderson CJ, et al. (2016) Response to comment on "Estimating the reproducibility of psychological science". *Science* 351(6277):1037.
48. Kahneman D (2014) A new etiquette for replication. *Soc Psychol* 45(4):299–311.
49. Reardon S (2016) A mouse's house may ruin experiments. Available at www.nature.com/news/a-mouse-s-house-may-ruin-experiments-1.19335. Accessed May 10, 2016.
50. Schaffer S (1989) Glass works: Newton's prisms and the uses of experiment. *The Uses of Experiment: Studies in the Natural Sciences*, eds Gooding D, Pinch T, Schaffer S (Cambridge Univ Press, Cambridge, UK), pp 67–104.
51. Bloom P (2016) Psychology's Replication Crisis Has a Silver Lining. Available at www.theatlantic.com/science/archive/2016/02/psychology-studies-replicate/468537/. Accessed May 10, 2016.
52. Roberts B (2015) The new rules of research. Available at https://pigee.wordpress.com/2015/09/17/the-new-rules-of-research/. Accessed May 10, 2016.
53. Srivastava S (2015) Moderator interpretations of the Reproducibility Project. Available at https://hardsci.wordpress.com/2015/09/02/moderator-interpretations-of-the-reproducibility-project/. Accessed May 10, 2016.
54. Koole SL, Lakens D (2012) Rewarding replications a sure and simple way to improve psychological science. *Perspect Psychol Sci* 7(6):608–614.
55. Simons DJ (2014) The value of direct replication. *Perspect Psychol Sci* 9(1):76–80.
56. Gilbert DT, King G, Pettigrew S, Wilson TD (2016) Comment on "Estimating the reproducibility of psychological science". *Science* 351(6277):1037.
57. Gilbert DT, King G, Pettigrew S, Wilson TD (2016) More on "Estimating the Reproducibility of Psychological Science". Available at projects.iq.harvard.edu/files/psychology-replications/files/gkpw_post_publication_response.pdf?platform=hootsuite. Accessed May 10, 2016.
58. Lakens D (2016) The statistical conclusions in Gilbert et al (2016) are completely invalid. Available at daniellakens.blogspot.nl/2016/03/the-statistical-conclusions-in-gilbert.html. Accessed May 10, 2016.
59. Srivastava S (2016) Evaluating a new critique of the Reproducibility Project. Available at https://hardsci.wordpress.com/2016/03/03/evaluating-a-new-critique-of-the-reproducibility-project/. Accessed May 10, 2016.
60. Etz A, Vandekerckhove J (2016) A Bayesian perspective on the reproducibility project: Psychology. *PLoS One* 11(2):e0149794.
61. Klein RA, et al. (2014) Investigating variation in replicability. *Soc Psychol* 45(3):142–152.
62. Schweinsberg M, et al. (2016) The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *J Exp Soc Psychol*, in press.
63. Eyal T, Liberman N, Trope Y (2008) Judging near and distant virtue and vice. *J Exp Soc Psychol* 44(4):1204–1209.
64. Schooler JW, Engstler-Schooler TY (1990) Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognit Psychol* 22(1):36–71.
65. Shih M, Pittinsky TL, Ambady N (1999) Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychol Sci* 10(1):80–83.
66. Williams LE, Bargh JA (2008) Experiencing physical warmth promotes interpersonal warmth. *Science* 322(5901):606–607.
67. Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 86(2):420–428.
68. Turk-Browne NB, Isola PJ, Scholl BJ, Treat TA (2008) Multidimensional visual statistical learning. *J Exp Psychol Learn Mem Cogn* 34(2):399–407.
69. Harmon-Jones E, Harmon-Jones C, Fearn M, Sigelman JD, Johnson P (2008) Left frontal cortical activation and spreading of alternatives: Tests of the action-based model of dissonance. *J Pers Soc Psychol* 94(1):1–15.
70. Colzato LS, et al. (2008) How does bilingualism improve executive control? A comparison of active and reactive inhibition mechanisms. *J Exp Psychol Learn Mem Cogn* 34(2):302–312.
71. Fiedler K (2008) The ultimate sampling dilemma in experience-based decision making. *J Exp Psychol Learn Mem Cogn* 34(1):186–203.
72. Purdie-Vaughns V, Steele CM, Davies PG, Ditlmann R, Crosby JR (2008) Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *J Pers Soc Psychol* 94(4):615–630.
73. Richard FD, Bond CF, Jr, Stokes-Zoota JJ (2003) One hundred years of social psychology quantitatively described. *Rev Gen Psychol* 7(4):331–363.
74. Bohannon J (2015) Reproducibility. Many psychology papers fail replication test. *Science* 349(6251):910–911.
75. Baker M (August 27th, 2015) Over half of psychology studies fail reproducibility test. Available at www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248. Accessed May 10, 2016.
76. Brandt MJ, et al. (2014) The replication recipe: What makes for a convincing replication? *J Exp Soc Psychol* 50:217–224.
77. Hajcak G, Foti D (2008) Errors are aversive: Defensive motivation and the error-related negativity. *Psychol Sci* 19(2):103–108.
78. Dai X, Wertenbroch K, Brendl CM (2008) The value heuristic in judgments of relative frequency. *Psychol Sci* 19(1):18–19.
79. Alter AL, Oppenheimer DM (2008) Effects of fluency on psychological distance and mental construal (or why New York is a large city, but New York is a civilized jungle). *Psychol Sci* 19(2):161–167.
80. Greene CS, Penrod NM, Williams SM, Moore JH (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One* 4(6):e5639.
81. Djulbegovic B, Hozo I (2014) Effect of initial conditions on reproducibility of scientific research. *Acta Inform Med* 22(3):156–159.
82. Cesario J (2014) Priming, replication, and the hardest science. *Perspect Psychol Sci* 9(1):40–48.
83. Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? *Behav Brain Sci* 33(2-3):61–83, discussion 83–135.
84. Leek JT, Peng RD (2015) Statistics. What is the question? *Science* 347(6228):1314–1315.
85. Pashler H, Harris CR (2012) Is the replicability crisis overblown? Three arguments examined. *Perspect Psychol Sci* 7(6):531–536.
86. Dreber A, et al. (2015) Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci USA* 112(50):15343–15347.
87. Monin B, et al. (2014) Commentaries and rejoinder on Klein et al. (2014). *Soc Psychol* 45(4):299–311.
88. Cacioppo JT, Berntson GG (1992) Social psychological contributions to the decade of the brain. Doctrine of multilevel analysis. *Am Psychol* 47(8):1019–1028.
89. Bevan W (1991) Contemporary psychology: A tour inside the onion. *Am Psychol* 46(5):475–483.
90. Sattler JM (1970) Racial "experimenter effects" in experimentation, testing, interviewing, and psychotherapy. *Psychol Bull* 73(2):137–160.
91. Anderson CA (1989) Temperature and aggression: Ubiquitous effects of heat on occurrence of human violence. *Psychol Bull* 106(1):74–96.
92. May CP, Hasher L, Stoltzfus ER (1993) Optimal time of day and the magnitude of age differences in memory. *Psychol Sci* 4(5):326–330.