American Psychological Association

Psychological Methods

© 2019 American Psychological Association
1082-989X/19/$12.00

2019, Vol. 24, No. 5, 571–575
http://dx.doi.org/10.1037/met0000223

COMMENT

Challenges and Suggestions for Defining Replication "Success" When Effects May Be Heterogeneous: Comment on Hedges and Schauer (2019)

Maya B. Mathur
Harvard T. H. Chan School of Public Health and Stanford University

Tyler J. VanderWeele
Harvard T. H. Chan School of Public Health

Psychological scientists are now trying to replicate published research from scratch to confirm the findings. In an increasingly widespread replication study design, each of several collaborating sites (such as universities) independently tries to replicate an original study, and the results are synthesized across sites. Hedges and Schauer (2019) proposed statistical analyses for these replication projects; their analyses focus on assessing the extent to which results differ across the replication sites, by testing for heterogeneity among a set of replication studies, while excluding the original study. We agree with their premises regarding the limitations of existing analysis methods and regarding the importance of accounting for heterogeneity among the replications. This objective may be interesting in its own right. However, we argue that by focusing only on whether the replication studies have similar effect sizes to one another, these analyses are not particularly appropriate for assessing whether the replications in fact support the scientific effect under investigation or for assessing the power of multisite replication projects. We reanalyze Hedges and Schauer's (2019) example dataset using alternative metrics of replication success that directly address these objectives. We reach a more optimistic conclusion regarding replication success than they did, illustrating that the alternative metrics can lead to quite different conclusions from those of Hedges and Schauer (2019).

Keywords: replication, meta-analysis, heterogeneity, replicability, reproducibility

Hedges and Schauer (2019) proposed statistical analyses for replication projects; these analyses focus on testing for heterogeneity among a set of replication studies, excluding the original study on which the replications were based. We commend their attention to heterogeneity in the replications and to conducting inference on the underlying true effects rather than the point estimates; we agree that existing analyses of replication projects have often overlooked these important considerations. They provide several useful tests of heterogeneity among the replications that are based on the meta-analytic *Q*-statistic and characterize their power to detect varying amounts of heterogeneity. We consider such tests to be quite useful for meta-analysis more generally, even outside the context of replication; substantial heterogeneity can, for example, suggest future directions regarding scientifically important moderators. As Hedges and Schauer (2019) demonstrate, heterogeneity tests require a large number of studies to achieve adequate power; for example, they show that at least 40 studies would be required to detect with 80% power heterogeneity with a magnitude two thirds as large the within-study error variances. Thus, if the scientific goal is to detect heterogeneity among the replications, nearly all multisite replications to date are indeed underpowered, as Hedges and Schauer (2019) conclude.

We agree with Hedges and Schauer's (2019) premises regarding the limitations of existing analyses conducted in multisite replication projects and regarding the importance of considering heterogeneity in these analyses. However, we also believe that their proposed statistical analyses based on assessing for heterogeneity in the effect sizes underlying the replications answer a question that may be interesting in its own right, but that is not directly relevant to evaluating the "success" or "failure" of a replication project in the sense that usually motivates these projects. As we will demonstrate, if their proposed heterogeneity analyses are interpreted as evidence regarding whether the original study successfully replicated, or regarding the extent to which the replica-

Maya B. Mathur, Department of Epidemiology, Harvard T. H. Chan School of Public Health, and Quantitative Sciences Unit, Stanford University; Tyler J. VanderWeele, Department of Epidemiology, Harvard T. H. Chan School of Public Health.

Maya B. Mathur and Tyler J. VanderWeele were supported by NIH Grant R01 CA222147. The funders had no role in the design, conduct, or reporting of this research. The datasets analyzed in this article were previously analyzed by Klein et al. (2014), Hedges and Schauer (2019), and Mathur et al. (in press). The replication metrics we apply here were first proposed in a preprint, Mathur and VanderWeele (2017). Reproducibility: All data and code required to reproduce the analyses are publicly available (https://osf.io/f6sh2/).

Correspondence concerning this article should be addressed to Maya B. Mathur, Quantitative Sciences Unit (c/o Inna Sayfer), Stanford University, 1070 Arastradero Road, Palo Alto, CA 94305. E-mail: mmathur@stanford.edu

tions support the scientific effect under investigation, the conclusions can be misleading.

Hedges and Schauer's (2019) proposed analyses are motivated by their definition of "replication" as the similarity of the true effect sizes in the replications to one another. Given a set of replication studies, such as Klein et al.'s (2014) 36 independent replications of an original study on the reverse gambler's fallacy, Hedges and Schauer (2019) proposed testing for heterogeneity among these replications, while excluding the original study because it is likely subject to publication bias. They consider null hypotheses postulating zero heterogeneity as well as null hypotheses postulating that there is less than a fixed, nonzero amount of heterogeneity. A large amount of heterogeneity is interpreted as failure to replicate. Yet we believe that this motivating definition of replication and the resulting statistical analyses do not align particularly well with the scientific questions of interest in replication studies. The goal is not to determine whether the replication studies are similar to one another, as in Hedges and Schauer's (2019) approach, but rather to determine whether they support the scientific effect under investigation and potentially also whether they are consistent with the findings of a "privileged" original study. Under Hedges and Schauer's (2019) framework, a set of replication studies all with exactly null point estimates would be considered a perfect replication "success," even if the original study hypothesized and supported a large positive effect. For example, we previously conducted a multisite replication (Mathur et al., in press) of an original study that reported a positive effect regarding a form of magical thinking (mean difference = 1.03; 95% CI [0.09, 1.97]; p = .03; Risen & Gilovich, 2008).[1] These 11 replications (n = 4,441 total) showed little apparent heterogeneity (Q = 1.85, p = 1.00), such that Hedges and Schauer's (2019) test would fail to reject the null hypothesis of exact replication. Yet meta-analyzing these replications[2] yields a very small, negative pooled point estimate with a large p value (Hedges' $g = -0.06$, 95% CI [−0.19, 0.07], p = .31) in contrast to the predicted positive effect. We believe it would be misleading to describe these replications as "successful" merely because they do not exhibit heterogeneity, when in fact they provide essentially no support for the investigated effect.

Conversely, in Hedges and Schauer's (2019) framework, a set of replication studies showing substantial heterogeneity, but in which all the point estimates are quite large, would be considered a replication "failure," even though such replications may strongly support the scientific effect under consideration. Indeed, this is exactly the case in Hedges and Schauer's (2019) example regarding the reverse gambler's fallacy. Hedges and Schauer's (2019) analysis indicates fairly high heterogeneity among these replications ($Q = 51.61$ vs. a critical value of $\chi^2_{35} = 49.80$), leading them to conclude that there is substantial heterogeneity, which certainly does seem to be the case. However, they also conclude from this a rejection of the null hypothesis of "exact replication." Consider, however, a random-effects meta-analysis of the 36 replication point estimates, yielding a pooled point estimate of Hedges' g = 0.61 (95% CI [0.54, 0.69]; p < 1e-05). This suggests that the average true effect size underlying the replications is large, with a narrow confidence interval that excludes small effect sizes. Additionally, this point estimate heuristically appears to be quite similar in magnitude to that of the original study (Hedges' g = 0.68; 95% CI [0.18, 1.19]; p = 8e-03; Oppenheimer & Monin, 2009). These basic measures provide strong evidence for the reverse gambler's fallacy, with an average effect size that is nearly as strong as that estimated in the original study. We would consider this an unambiguously successful replication.

However, these basic reanalyses do not characterize heterogeneity among the replications; they focus only the estimated mean of the true effects underlying the replications' point estimates. Might characterizing the full distribution of the apparently heterogeneous true effects underlying the replications lead us to Hedges and Schauer's (2019) more pessimistic conclusion regarding replication success? To investigate, we meta-analytically estimated the proportion of these effects that were stronger than several choices of thresholds representing a minimum scientifically meaningful effect size. We have demonstrated elsewhere that this metric, termed $\hat{P}_{>q}$, better characterizes evidence strength in a potentially heterogeneous population of effect sizes than the pooled point estimate alone (Mathur & VanderWeele, 2017, 2019). Furthermore, this metric fulfills Hedges and Schauer's (2019) apt stipulation that statistical analyses of replication studies focus on effect sizes rather than p values and that they account for statistical error and heterogeneity. For our reanalysis, we used the R package Replicate to estimate the proportion of true effects in the replications (excluding the original study) above increasingly stringent Hedges' g values of 0, 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6. In this apparently heterogeneous population of true effects (estimated standard deviation $[\hat{\tau}]$ = 0.09; 95% CI [0.00, 0.16]; estimated percent of between-study variance due to heterogeneity $[\hat{I}^2]$ = 24.71), we estimated that an overwhelming majority of true effects are stronger than even quite stringent thresholds (see Table 1). For example, if we consider only effects surpassing a fairly large effect size of Hedges' g = 0.50 to be of scientifically meaningful size, we nevertheless estimate that 88% (95% CI [56%, 100%]) of the true effects in the various studies meet this stringent criterion. Therefore, a principled meta-analytic examination of the distribution of the true effects in these replications again leads us to the opposite conclusion from that of Hedges and Schauer (2019): We conclude that these replications, despite their heterogeneity, were

Table 1
Estimated Proportion of True Effect Sizes (Hedges' g) in Klein et al.'s (2014) Replications Surpassing Various Thresholds Defining a Scientifically Meaningful Effect Size

Threshold	$\hat{P}_{>q}$ [95% CI]
.00	1.00 [1.00, 1.00]
.10	1.00 [.99, 1.00]
.20	1.00 [.96, 1.00]
.30	1.00 [.90, 1.00]
.40	.99 [.78, 1.00]
.50	.88 [.56, 1.00]
.60	.55 [.24, .86]

[1] Approximate effect sizes were recomputed from rounded values in the original article.

[2] For ease of presentation, these reanalyses have been simplified somewhat compared with the models originally fit by Mathur et al. (in press), but yielded qualitatively similar results.

remarkably successful in that they provide strong evidence for the reverse gambler’s fallacy.

Hedges and Schauer (2019) avoid comparing the original study to the replications because the former is likely subject to publication bias and because “comparing an initial study with an aggregate finding from replications may not address lack of agreement among the replications.” However, it is possible to directly compare the original study with the replications in a manner that accounts correctly for heterogeneity and statistical error in both the original study and the replications. We might, for example, consider the original study to be “consistent” with the replications if it is generated from the same underlying distribution as the replications; that is, its true effect size comes from the same distribution as those of the replications, again allowing for the possibility that this distribution may be heterogeneous. We showed elsewhere how to calculate the probability that, if indeed the original is consistent with the replications in this sense, its estimate would be as extreme or more extreme than it actually was (a metric termed P_{orig}; Mathur & VanderWeele, 2017). This metric can be interpreted similarly to a *p* value in that a small value of P_{orig} would indicate strong evidence that the original study is inconsistent with the replications (due, example, to publication bias), whereas a large value would suggest relatively good consistency.[3] For the reverse gambler’s fallacy example, we used the R package Replicate to estimate $P_{orig} = 0.80$, indicating good consistency between the replications and the original study. Along with the analyses presented above, this suggests a successful replication of the original gambler’s fallacy study.

Regarding statistical power, Hedges and Schauer (2019) report that the Many Labs project was underpowered to detect interreplication heterogeneity, and we agree with the statistical validity of their power analysis. However, these power calculations are again based on a hypothesis test that may be interesting in its own right to characterize heterogeneity, but that does not address the scientific goals for which these replication projects are in fact designed and powered. In contrast, our reanalysis estimated a precise confidence interval for the pooled point estimate (i.e., 95% CI [0.54, 0.69]) and reasonably precise confidence intervals for $\hat{P}_{>q}$, particularly for smaller thresholds. This demonstrates that a multisite replication project may be adequately powered to estimate the strength of evidence for the effect under investigation while accounting for interreplication heterogeneity, even if its statistical power to test for interreplication heterogeneity itself is low. Statistical power to detect inconsistency via P_{orig} may be limited if the original study or the replications were low-powered or few in number. If the original study was low-powered, but the replications were well-powered, then $\hat{P}_{>q}$ may still be estimated precisely because it is not a function of the original study’s point estimate. We are therefore optimistic that when considering evidence for the gambler’s fallacy effect rather than inter-replication heterogeneity as an end in itself, this Many Labs replication was adequately powered, whether focus lies on the pooled point estimate or on metrics that additionally account for heterogeneity.

In fact, depending on the design of the replication studies, one might even argue that heterogeneous replication studies that strongly support the original effect under investigation (evidenced, example, by a large, statistically precise point estimate, along with a large value of $\hat{P}_{>q}$) may provide *more* persuasive evidence for the effect under investigation than homogeneous replication studies with a similar pooled point estimate. For example, if the replication studies are heterogeneous because they recruit demographically or ideologically diverse subjects, they may support a robust effect that generalizes to subjects unlike those recruited in the original study (Jones, 2010). Alternatively, others have advocated conducting “conceptual replications” that assess the same theory as the original study, but using different operationalizations (Crandall & Sherman, 2016; Lynch, Bradlow, Huber, & Lehmann, 2015; Monin et al., 2014). If an analysis of such replications suggests that these different operationalizations overall strongly support the effect under investigation, despite their intentional heterogeneity, this may provide compelling evidence for conceptual robustness in addition to direct replicability. Yet in both of these contexts, Hedges and Schauer’s (2019) proposed analyses will again reach exactly the opposite conclusion, namely that the replication studies failed to replicate simply because they are heterogeneous.

Table 2
Differences in Interpretation Between Three Measures of Replication Success

Metric	Consistency between original and replications	Evidence strength for large effect sizes in replications	Heterogeneity among replications
P_{orig}	✓		
$\hat{P}_{>q}$		✓	
Hedges & Schauer’s *Q*			✓

To further clarify the distinction between operational definitions of replication “success” under Hedges and Schauer’s (2019) versus our proposed methods, Table 2 summarizes the questions each method is best suited to address. P_{orig} helps assess whether the replications were “successful” in the sense that their results are similar to those of the original study. In contrast, $\hat{P}_{>q}$ helps assess whether the replications were “successful” in the sense of providing evidence for the effect under investigation, regardless of the results of the original study.[4] Last, Hedges and Schauer’s (2019) test assesses for heterogeneity among the replications. To illustrate how each metric would characterize replication “success” for different observed patterns of replication data, Figure 1 shows histograms representing 25 simulated replication study point esti-

[3] More formally, P_{orig} can be interpreted as a *p*-value for testing the null hypothesis H_0: $\hat{\theta}_{orig} \sim N(\mu, \tau^2)$ versus H_A: not $\hat{\theta}_{orig} \sim N(\mu, \tau^2)$, where $\hat{\theta}_{orig}$ is the point estimate of the original study and μ and τ^2 are the unknown parameters of the distribution of true effects. The test assumes that the aggregated replication studies provide unbiased estimates of μ and τ^2 (Mathur & VanderWeele, 2017). Like a *p*-value, P_{orig} is best interpreted as a continuous measure of evidence for inconsistency rather than as an arbitrarily dichotomized indicator of “statistical significance” (e.g., Wasserstein & Lazar, 2016). We therefore discourage describing P_{orig} as “significant” or “nonsignificant” based on an assessment of whether it is smaller than 0.05, for example.

[4] We further discuss the distinction between these two metrics in the applied example in Mathur and VanderWeele (2017), as well as the statistical assumptions required for each. For example, these methods should be applied only when there are at least 10 replication studies and when the replication studies are not subject to publication bias.

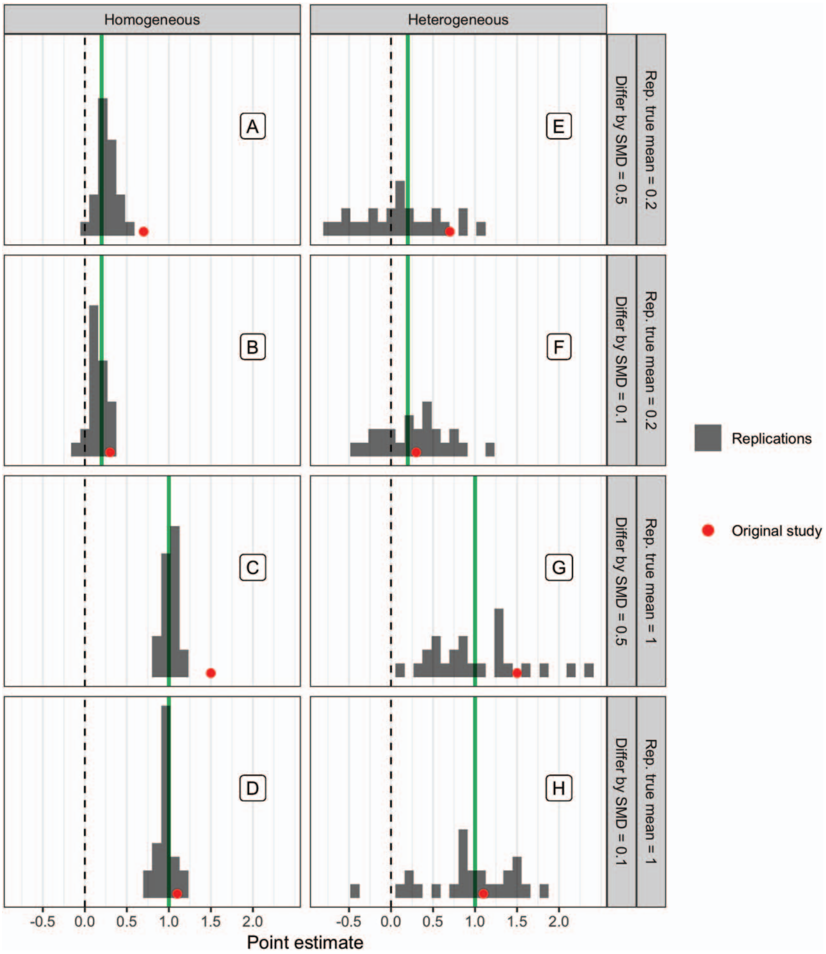

Figure 1. Simulated replication point estimates (histograms) arising from a homogeneous (variance = 0) or heterogeneous (variance = 0.25) normal distribution of true effects with mean of 0.2 or 1.0 (solid green lines). The original study point estimates (red points) differ from the true mean of the replications by SMD = 0.5 or SMD = 0.1. All studies have standard errors of 0.1. Dashed black lines indicate the null. See the online article for the color version of this figure.

mates arising from a homogeneous (variance = 0) or heterogeneous[5] (variance = 0.25) distribution of true effects with a mean of 0.2 or 1.0. The original study point estimates differ from the true mean of the replications by a standardized mean difference (SMD) of 0.5 or 0.1. All studies have standard errors of 0.1. Table 3 shows the results of applying Hedges and Schauer's (2019) test as well as P_{orig} and $\hat{P}_{>q}$ with $q = 0.5$ in each scenario. Hedges and Schauer's (2019) test against the null hypothesis of no heterogeneity correctly identifies heterogeneity in the four scenarios indeed generated with heterogeneous replications ($p <$ 1e-05, Scenarios E–H in Figure 1 and Table 3) and does not provide strong evidence of heterogeneity for the scenarios indeed generated with homogeneous replications (Scenarios A–D). Thus, Scenarios E–H would also be considered replication "failures" in Hedges and Schauer's (2019) framework, while Scenarios A–D would be considered replication "successes." Although these results accurately assess heterogeneity among the replications, they do not distinguish between scenarios in which the original study point estimate is considerably larger than the mean of the replication distribution

Table 3
Comparison of Metrics of Replication Success Across Scenarios Shown in Figure 1

Scenario	P_{orig}	$\hat{P}_{>q}$	CI for $\hat{P}_{>q}$	Q	p-value for Q
A	1e-04	.00	[.00, .03]	33.53	.09
B	.17	.00	[.00, .00]	25.93	.36
C	<1e-05	1.00	—	13.83	.95
D	.21	1.00	[1.00, 1.00]	26.17	.34
E	.22	.20	[.07, .33]	564.61	<1e-05
F	.96	.28	[.13, .43]	383.63	<1e-05
G	.41	.82	[.70, .95]	788.09	<1e-05
H	.75	.79	[.65, .92]	705.43	<1e-05

Note. "—" = CI could not be estimated because the heterogeneity estimate was exactly 0.

[5] The amount of heterogeneity is exaggerated here in order to enhance the visual distinction for illustrative purposes; it is not meant to represent the amount of heterogeneity typically seen in multisite replications.

versus close to the mean (e.g., Scenarios A vs. B). In contrast, P_{orig} does help distinguish between these scenarios (P_{orig} = 1e-04 vs. P_{orig} = 0.17, respectively.) Additionally, Hedges and Schauer's (2019) test does not distinguish between scenarios in which the replication effect sizes are small versus large (e.g., Scenarios E vs. G). In contrast, $\hat{P}_{>q}$ estimates a considerably higher proportion of large effect sizes in Scenario G ($\hat{P}_{>q} =$ 0.82) than in Scenario E ($\hat{P}_{>q} =$ 0.20).

In closing, we emphasize our agreement with Hedges and Schauer's (2019) recommendation to analyze replication projects using metrics that properly account for heterogeneity. We further echo their concerns about commonly reported existing metrics, such as those focusing exclusively on the "statistical significance" of replication results, and we certainly hope that the empirical replication literature will take note. But, critically, we disagree with Hedges and Schauer's (2019) definition of replication success and with their resulting conclusion that the existing multisite replication studies are underpowered to evaluate replication success. Their power analysis for detecting heterogeneity is correct, but we have argued that this is often not what is of primary interest in assessing replications. Metrics for replication success must directly address whether the replications provide evidence for the effect under investigation, not whether the replications are similar to one another; it is here that Hedges and Schauer's (2019) analyses are not adequate when used as measures of replication success rather than to address distinct questions regarding heterogeneity among the replications (e.g., to investigate whether there may be important unknown moderators of the effect under investigation). In our reanalyses, we have demonstrated some alternative metrics that we believe address both considerations and build upon previous methods that apply when there is one replication per original study (e.g., Anderson & Maxwell, 2016; Etz & Vandekerckhove, 2016; Gilbert, King, Pettigrew, & Wilson, 2016; Open Science Collaboration, 2015; van Aert & Van Assen, 2017). We also look forward to further developments in the literature.

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21,* 1–12.

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

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE, 11,* e0149794.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037–1103.

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods, 24,* 557–570. http://dx.doi.org/10.1037/met0000189

Jones, D. (2010). A WEIRD view of human nature skews psychologists' studies. *Science, 328,* 1627.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology, 45,* 142–152.

Lynch, J. G., Jr., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing, 32,* 333–342.

Mathur, M. B., Bart-Plange, D. J., Aczel, B., Bernstein, M. H., Ciunci, A., Ebersole, C. R., . . . Frank, M. C. (in press). Many Labs 5: Registered multisite replication of tempting-fate effects in Risen & Gilovich (2008). *Advances in Methods and Practices in Psychological Science*.

Mathur, M. B., & VanderWeele, T. J. (2017). *New statistical metrics for multisite replication projects*. Retrieved from https://osf.io/w89s5/

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine, 38,* 1336–1342.

Monin, B., Oppenheimer, D. M., Ferguson, M. J., Carter, T. J., Hassin, R. R., Crisp, R. J., . . . Klein, R. A. (2014). Commentaries and rejoinder on Klein et al. (2014)>. *Social Psychology, 45,* 299–311.

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

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

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology, 95,* 293–307.

van Aert, R. C., & Van Assen, M. A. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS ONE, 12,* e0175302.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70,* 129–133.

Received November 16, 2018
Revision received March 18, 2019
Accepted March 28, 2019 ■