Big secrets do not necessarily cause hills to appear steeper

Etienne P. LeBel

Montclair State University

Christopher J. Wilbur

University of Wisconsin – Colleges

In press, *Psychonomic Bulletin & Review*

Corresponding author:
Etienne P. LeBel
etienne.lebel@gmail.com

Word count: 3,602 (Excl. abstract, title page, acknowledgments, figures, figure captions, and tables)

Abstract

Slepian, Masicampo, Toosi, and Ambady (2012, Study 1) found that individuals recalling and writing about a big, meaningful secret judged a pictured hill as steeper than those who recalled and wrote about a small, inconsequential secret (with estimates unrelated to physical effort unaffected). From an embodied cognition perspective, this result was interpreted as suggesting that important secrets weigh people down. Answering to mounting calls for the crucial need of independent direct replications of published findings to ensure the self-correcting nature of our science, we sought to corroborate Slepian et al.'s finding in two extremely high-powered, pre-registered studies that were very faithful to all procedural and methodological details of the original study (i.e., same cover story, study title, manipulation, measures, item order, scale anchors, task instructions, sampling frame, population, and statistical analyses). In both samples, we were unsuccessful in replicating the target finding. Though Slepian et al. reported three other studies supporting the secret burdensomeness phenomenon, we advise that these three other findings need to be independently corroborated before the general phenomenon informs theory or health interventions.

Key words: embodied cognition, secrecy, concealment of secrets, independent direct replication

Word count: 176

Big secrets do not necessarily cause hills to appear steeper

In recent years, psychological science has experienced a rapidly growing interest in embodied cognition (e.g., Schnall, Benton, & Harvey, 2008; Vess, 2012, see Landau, Meier, & Keefer, 2010, for a review). According to the embodied cognition perspective, the body and the mind are inextricably linked in that bodily states influence mental processes and vice versa (Barsalou, 2008; Landau et al., 2010). For example, thinking about past social exclusions has been shown to cause people to feel physically colder (IJzerman & Semin, 2009) and holding a warm coffee cup caused individuals to judge a target as more interpersonally warm (Williams & Bargh, 2008). Given that the embodied cognition perspective has been applied to several classic diverse domains of psychological inquiry, including romantic attachment (Vess, 2012), moral judgment (e.g., Schnall et al., 2008), and visual perception (e.g., Cole, Balcetis, & Zhang, 2013), it offers a potentially parsimonious account for explaining myriad forms of human thought and behavior. A fundamental scientific principle, however, is that particular findings must be shown to be replicable before they become accepted as genuine contributions to human knowledge. Indeed, there are mounting calls for conducting independent direct replications to ensure the self-correcting nature of our science (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Ioannidis, 2012; Koole & Lakens, 2012; Makel, Plucker, & Hegarty, 2012; Neuliep & Crandall, 1990; Nosek, Spies, & Motyl, 2012; Pashler & Wagenmakers, 2012; Schimmack, 2012). In this spirit, we sought to replicate a potentially important recent finding by Slepian, Masicampo, Toosi, and Ambady (2012) on the embodiment of secrets.

Guided by an embodied cognition perspective, Slepian et al. (2012) reasoned that because secrets mentally tax the secret bearer, they might also be experienced as physically taxing. Given that being burdened by physical weight has previously been shown to influence perceptions related to physical effort (Proffitt, 2006), Slepian et al. hypothesized that harboring important secrets would result in perceiving the physical environment as more demanding and limit physical forms of helping.

In their first study, Slepian et al. found that participants recalling and writing about a big, meaningful secret judged a pictured hill as steeper than those who recalled and wrote about a small, inconsequential secret. Estimates irrelevant to physical taxation (e.g., durability of a water bottle) did not differ between the two groups. Presumably those “weighed down” by a large secret judged the physical terrain as particularly arduous, as if they were encumbered by a heavy backpack (Proffitt, 2006). Three subsequent conceptual replications provided additional evidence supporting this general idea, using different operationalizations of the independent and dependent variables. In Study 2, Slepian et al. found that individuals recalling a big secret overthrew a beanbag at a target more so than individuals recalling a small secret, presumably because they perceived the distance to the target as greater. Slepian et al. further showed that individuals concealing important secrets (compared to trivial secrets) perceived physical tasks as more effortful (Study 3) and engaged in less prosocial behavior involving physical tasks (Study 4).

Slepian et al.’s (2012) findings offer substantial potential if they prove to be robust. For example, they could have important applied counseling implications for mitigating possible negative health consequences in individuals who are concealing weighty information such as sexual orientation. Given the theoretical and applied promise of these findings, and recent calls for the dire need of independent direct replications (Makel et al., 2012), we attempted to corroborate Slepian et al.’s Study 1 results.

In two large samples, we attempted to replicate Slepian et al.’s (2012) Study 1 finding using the exact same procedures, manipulation, measures, sampling type, and population. We contacted the corresponding author (M. Slepian) to acquire as many of the procedural and methodological details as possible, used large sample sizes to ensure high statistical power[1], and pre-registered the methods and planned statistical analyses prior to data collection for full transparency (LeBel et al., in press; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).[2]

For our first attempt, M. Slepian provided the title of the study used to recruit participants, cover story, instructions, exact wording and nature of the secret manipulation (big vs. small), stimuli used for the dependent variables (photos of the control items – a table, a water bottle, and a park – and the critical hill steepness item used in the original study), and order of the dependent variables. We used the same sample type (online via Amazon's Mechanical Turk) and sampling frame (adults ranging from 18 to 75 years of age). Slepian could not provide the exact compensation amount used in the original study, so we decided on a \$0.25 USD compensation given that Slepian stated that they tend to pay between \$0.05 USD and \$0.25 USD for online Mechanical Turk studies in their lab.

For our second attempt, we sought to further maximize the precision of our direct replication attempt by inviting M. Slepian to review all of the procedural and methodological details used in our first attempt. M. Slepian graciously agreed and in doing so clarified the exact wording of the instructions for introducing the dependent variables, provided the exact description of the study as advertised on Mechanical Turk, and also indicated that the answer boxes to the park temperature control item and hill steepness critical item were below the photos (rather than above as in our first attempt). We implemented these minor procedural

---

[1]Power analyses indicated that a sample size of 89 would be needed to achieve a power level of .95 (power estimated using G-Power 3.1; Faul, Erdfelder, Buchner, & Lang, 2009), based on the effect size of the critical comparison between big and small secret conditions in the original study (d = .78). Given the inexpensive nature of the online sampling type used by the original authors we decided to aim for N=200 in both studies. Due to the unpredictable nature of online sign-up patterns, however, the number of complete data points was higher for our first sample and much lower for our second sample. In the latter case, sign-ups unexpectedly stopped after one week, and after almost two weeks of inactivity we decided to halt the study rather than increasing the compensation which could have introduced a self-selection confound.

[2]Pre-registration involves specifying methodological and analytical plans in a frozen time-stamped document prior to data collection so that stringent confirmatory tests of the relevant hypotheses can be achieved (Wagenmakers et al., 2012). Exact details of both replication attempts can be confirmed by cross-referencing the pre-registered replication protocols for replication attempt #1 and #2 available at http://bit.ly/16MSSx8 and http://bit.ly/11ngkZK, respectively.

changes for our second replication attempt. Also, M. Slepian stated that our consent form mentioning that participants' de-identified data may be shared with other researchers for re-analysis could have influenced the results given the manipulation involves revealing very personal information. It is important to note, however, that participants were explicitly informed – following Slepian et al. – that all information they provided would remain completely anonymous. Nonetheless, to address this possible concern, we moved the consent for data sharing in the post-experiment debriefing and significantly shortened the consent form (which was okayed by our ethics board). A revised document summarizing all of these minor changes was re-sent to M. Slepian and subsequently endorsed by him prior to commencing data collection for our second replication attempt.

We analyzed the data following the exact same analytic approach used by Slepian et al. (2012).[3] We first transformed the four dependent measures into standardized scores and averaged the three control items (judgments of a table's sturdiness, a water bottle's durability, and a park's temperature) to create an index of control numerical estimation. A 2 (condition: big vs. small secret) × 2 (measure type: hill steepness vs. control estimates) mixed-model ANOVA was then executed, with condition as a between-subjects factor and measure type as a within-subjects factor. Follow-up *t*-tests were used to test the critical difference in hill steepness estimates across secrecy conditions. Following Slepian et al., we excluded participants who provided invalid answers to the open-ended items (i.e., not providing a numerical estimate for the park control item; not providing an estimate between 0 and 90 degrees for the hill steepness item), resulting in two exclusions in our second sample.

[3] In the spirit of open science practices, de-identified raw data and syntax files for both of our replication attempts are available at http://openscienceframework.org/project/w6kV5/ and http://openscienceframework.org/project/EUZwH/.

Table 1. Interaction effects and critical mean comparisons of hill steepness estimates across conditions in Slepian et al. (2012, Study 1) and current studies

Study	N	Interaction Effect *F*	*p*	Effect Size (*r*)	Mean Comparisons of Hill Steepness Estimates: Big	Small	*t*	*p*	Effect size (*d*)	C.I.	Power
Slepian et al. (2012, Study 1)	40	13.99	.001	.52	46.05° (16.40°)	32.90° (17.98°)	2.42	.02	.784	[.12, 1.40]	67.5%
Current studies											
Sample 1	240	.834	.362	.06	37.79° (15.21°)	35.21° (14.23°)	1.35	.177	.176	[-.08, .43]	99.9%
Sample 2	90	3.34	.071	.19	39.33° (15.10°)	44.76° (19.12°)	-1.50	.139	-.319	[-.73, .10]	95.7%
Overall	330	.078	.780	.02	38.22° (15.15°)	37.77° (16.19°)	.264	.795	.029	[-.18, .25]	-

Note. Standard deviations in parentheses. C.I. = 95% confidence interval of the effect size. Overall effects were calculated based on combined samples. *Power* is the probability of detecting an effect as large (or larger) than the one reported by Slepian et al.

As shown in Table 1, we did not replicate Slepian et al.'s Study 1 finding in either sample, with interactions in both samples not statistically significant.[4] The interaction in our second sample was marginally significant ($p < .07$), however, the pattern was in the opposite direction from Slepian et al.'s original finding such that hill steepness estimations in our sample were numerically *smaller* (rather than *larger*) in the big secret compared to small secret condition (see Figure 1, panel C).

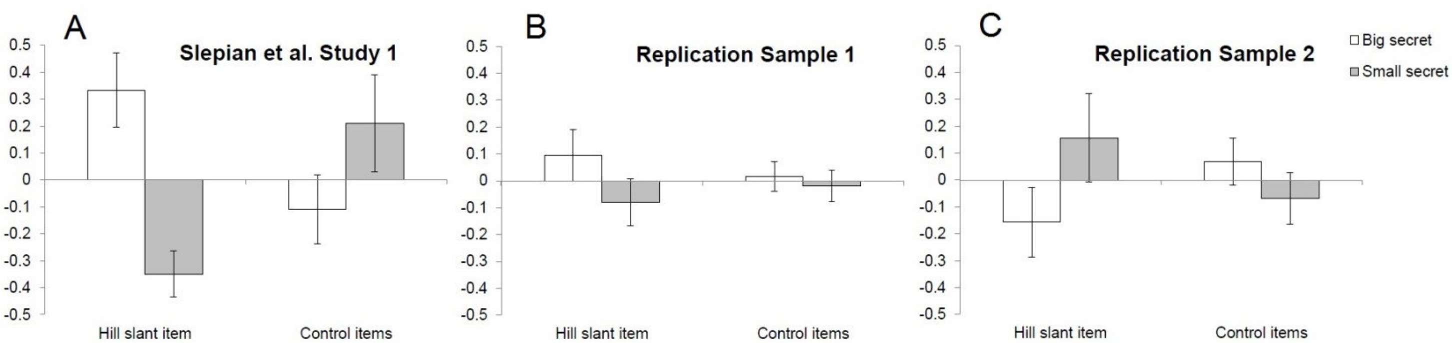

Figure 1. Means of critical hill steepness estimates and control items across conditions from original study (panel A) and in our first and second replication attempts (panel B and C, respectively). Error bars denote standard error of the mean.

[4]One participant in sample 2 was excluded from the analyses because this participant indicated s/he had previously participated in a study of the same name. Including this participant revealed the same pattern of results (non-statistically-significant interaction, $F(1, 89) = 2.69$, $p < .11$).

Additional clarity can be gained in interpreting our results via a Bayesian analysis, which quantifies the strength of evidence the data provide for or against the null hypothesis relative to the alternative hypothesis (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Employing a Bayes Factor (BF) test for two-group designs using a noninformative Jeffrey-Zellner-Siow prior (Rouder, Speckman, Sun, Morey, & Iverson, 2009) revealed a BF of .38 for Slepian et al.'s (2012) Study 1 hill steepness contrast and a BF of 11.13 in our combined sample (N = 330).[5] This indicates that our data provide 11 times more evidence for the null than the alternative hypothesis whereas Slepian et al.'s data provide only about 2.6 times (inverse of .38) more evidence for the alternative than the null hypothesis. In other words, our replication results provide much more compelling evidence in favor of the null than Slepian et al.'s evidence in favor of the burdensomeness of secrets alternative hypothesis.

Findings from our replication attempts are difficult to reconcile with Slepian et al.'s (2012) Study 1 results for several reasons. Our samples were extremely high-powered and were very faithful to all procedural and methodological details of the original study (i.e., same study title, cover story, manipulation, measures, item order, scale anchors, task instructions, sampling frame, and population). Both replication attempts were also pre-registered, ruling out undisclosed flexibility in design specifications and/or analyses being responsible for our results (Wagenmakers et al., 2012).

One potentially consequential difference between our replication attempts and the original study involved the use of a consent form in our first sample mentioning possible sharing of participants' data, which could have influenced the secrecy manipulation. This concern can be ruled out, however, given that absolutely no mention of data sharing was made in our second replication attempt, which also did not yield the expected pattern of results; in fact, it is noteworthy that the results in our improved second attempt were more discrepant to Slepian et

[5]These analyses were executed using Rouder et al.'s (2009) online calculator (http://pcl.missouri.edu/bf-two-sample) using the default scaling factor of r = 1 and relevant *t*-values and *n*s (i.e., n1=20, n2=20, and *t*=2.42 for Slepian et al.'s (2012) data and n1=162, n2=168, and *t*=.264 for our combined data).

al.'s finding than our first attempt (hill steepness mean difference was in opposite direction). The only other known difference in our replication attempts involved participant demographics. Slepian et al.'s Study 1 sample involved 65% females (mean age of 32.0 years) whereas our first sample involved only 36% females (mean age of 28.9 years). Hence, perhaps this sex composition difference contributed to our different results. However, analyses involving strictly females (N=87, power = 94%) in our first sample also failed to replicate Slepian et al.'s pattern (interaction $F<1$; and interestingly, hill steepness means in the opposite direction as in our second sample). Furthermore, our second sample (68% females; mean age of 35.9 years) closely matched the original's study sex breakdown, but still did not yield the expected result. Taken together, the consent form and demographic differences cannot account for the discrepant findings observed in our replication attempts.

Another potential concern is that our failed replication attempts were due to random or careless responding by the online participants, which of course would preclude the possibility of observing *any* statistically significant patterns. One way to rule out such concern is by verifying the reliability of the measures. This is not possible, however, because the main dependent variable is a 1-item measure and the control items are completely unrelated.[6] However, we were able to gauge the reliability of responses on a health questionnaire (Bhalla & Proffit, 1999) which was assessed after the main dependent variables for exploratory reasons (see pre-registered replication protocols for details). Reliability estimates for responses on this questionnaire were high (Cronbach's alpha were $\alpha = .75$ and $\alpha = .80$ in our first and second samples, respectively), which rules out the concern that our replication samples simply reflected random responding given that several questionnaire items were keyed negatively . In addition, we examined participants' completion times. Across our two samples, participants took an average of 4.83 minutes (*SD* = 2.83, median = 4.06) to complete the studies, with no participant

[6]Curiously, Slepian et al reported a reliability estimate of $r = .55$ (Spearman-Brown formula) for the control items in Study 1. This is puzzling given one would not expect these items to inter-correlate given they assess completely unrelated constructs (i.e., sturdiness of a table, durability of a water bottle, and temperature of a park).

taking less than one minute. Such completion times are inconsistent with the idea that participants responded carelessly, given the very brief nature of the study (simple manipulation and approximately 25 Likert-type questions). Furthermore, excluding participants who completed the study in less than two minutes (N = 6) or three minutes (N = 74) revealed identical pattern of results (interaction *F*s < 1). Taken together, these additional analyses rule out the alternative explanation that our replication failures were simply due to random or careless responding.

In conclusion, despite considerable effort and care to duplicate all procedural and methodological parameters of the original study, we failed to corroborate – in two high-powered replication attempts – Slepian et al.'s (2012) finding that big secrets cause hills to appear steeper.[7] That being said, our results cannot speak to the robustness of Slepian et al.'s three other particular findings, which used different operationalizations of the independent and dependent variables. For instance, our results do not speak to whether Slepian et al.'s Study 2 finding would independently replicate, whereby individuals recalling a big (compared to small) secret were more likely to overthrow a beanbag at a target. The robustness of these other particular findings is unknown at this time given that no (known) attempts have been made to independently confirm these other findings. Though it is generally understood that it is *sets* of particular findings *taken together* that provide evidence in support of a general idea, it is of course necessary that each particular finding in these sets is replicable under the conditions specified by the original researchers (Pashler & Harris, 2012).

These considerations are consistent with recent realizations regarding the severe limitations of an over-focus on *conceptual replications* (LeBel & Peters, 2011). That is, the exclusive publishing of *conceptual replications* – where researchers seek to replicate a finding

[7]Our results can also be considered failed replication attempts according to a new and improved standard proposed by Simonsohn (2013), whereby a replication attempt should be considered a failure if the effect observed in the replication attempt is too small to have been detected by the original study. Details of the analyses leading to such conclusion are available from the first author.

using *different* manipulations or measures – can lead to gross mischaracterization of the reality of psychological phenomena because particular findings never stand a chance of being disconfirmed (Pashler & Harris, 2012; Popper, 1959). Consequently, failed conceptual replications of a particular finding can justifiably be ignored (because the method was intentionally changed) leading to collections of conceptually-related findings for which the robustness of each particular finding is completely unknown. Taken together, our results lead us to advise researchers to await independent corroboration of Slepian et al.'s (2012) three other findings before the general secret burdensomeness phenomenon informs theory or guides health interventions for individuals concealing weighty secrets.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A, Fiedler, K, …, Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108-119.

Bakker, M., van Dijk, A, & Wicherts, J. M., (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Bhalla, M., & Proffit, D. R. (1999). Visual-motor recalibration in geographical slant perception. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1076-1096.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cole, S., Balcetis, E., & Zhang, S. (2013). Visual perception and regulatory conflict: Motivation and physiology influence distance perception. *Journal of Experimental Psychology: General, 142*, 18-22.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

IJzerman, H., & Semin, G. R. (2009). The thermometer of social relations: Mapping social

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives in Psychological Science, 7,* 645–654.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*, 608-614.

Landau, M. J., Meier, B. P., & Keefer, L. A. (2010). A metaphor-enriched social cognition. *Psychological Bulletin, 136,* 1045-1067.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*, 371-379.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (in press). PsychDisclosure.org: Grassroot support for reforming reporting standards in psychology. *Perspectives on Psychological Science.*

Makel, M. C., Plucker, J. A., & Hagerty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537-542.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*, 85–90.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528-530.

Popper, K. R. (1959). *The logic of scientific discovery.* Oxford, UK: Basic Books.

Proffitt, D. R. (2006). Embodied perception and the economy of action. *Perspectives on Psychological Science, 1*, 110–122.

proximity on temperature. *Psychological Science*, *20*, 1214–1220.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551-566.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science, 19*, 1219-1222.

Simonsohn, U. (2013). Evaluating replication results. Unpublished manuscript available at SSRN: http://ssrn.com/abstract=2259879 or http://dx.doi.org/10.2139/ssrn.2259879

Slepian, M. L., Masicampo, E. J., Toosi, N. R., & Ambady, N. (2012). The physical burdens of secrecy. *Journal of Experimental Psychology: General, 141,* 619-624.

Vess, M. (2012). Warm thoughts: Attachment anxiety and sensitivity to temperature cues. *Psychological Science, 23*, 472-474.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426-432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 627-633

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*, 606–607.

Authors' Note

We would like to thank Michael Slepian for his cooperation in providing study materials and procedural and methodological details. We also thank Yang Ye for preparing online study materials and for comments on an earlier version of this manuscript. This research was supported by a Social Science and Humanities Research Council (SSHRC) post-doctoral fellowship to the first author.