aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(5) 579–584

DOI: 10.1177/1745691614547878
pps.sagepub.com
SAGE

Turning the Lens of Science on Itself: Verbal Overshadowing, Replication, and Metascience

Jonathan W. Schooler
University of California, Santa Barbara

Abstract

This issue of *Perspectives on Psychological Science* reports an unprecedented replication effort entailing numerous independent laboratories conducting two versions of the verbal overshadowing paradigm (Schooler & Engstler-Schooler, 1990) using different timing intervals. The results (Alogna et al., 2014, this issue) provide unequivocal support for the existence of verbal overshadowing—the finding that describing a previously seen face can impair its subsequent recognition—while simultaneously revealing a number of factors that may have contributed to challenges in replicating verbal overshadowing in the past. In this commentary, I review my participation in this process and consider the implications of the results of this replication effort for verbal overshadowing, the decline effect, and the general goal of metascience: turning the lens of science onto itself.

Keywords

decline effect, eyewitness memory, open science, replication, verbal overshadowing

When *Perspectives in Psychological Science* asked me whether I would be interested in helping them develop a multisite replication of the verbal overshadowing (VO) paradigm (Alogna et al., 2014, this issue), it seemed like a win–win situation, as any outcome would support an effect that I had previously endorsed. A replication would be a win for the *VO effect* (Schooler & Engstler-Schooler, 1990)—the sometimes difficult to replicate (Meissner & Brigham, 2001; Schooler, 2011) and sometimes disputed (Francis, 2012) finding that describing a previously seen face can interfere with subsequent recognition of it. A failure to replicate would be a win for the decline effect (Jennions & Møller, 2002; Lehrer, 2010; Schooler, 2011)—the notion that science routinely observes effect sizes decrease over repeated replications for reasons that are still not well understood. However, during the replication process, I learned that when one encounters what seems to be a win–win situation, beware of Door #3! Near the completion of the initial replication study, it emerged that the research protocol that I had vetted included timing intervals that deviated from the original protocol in an important respect. However, this unexpected negative turn of events itself took a positive spin. The deviation in the initial protocol led to a final replication product that was far more informative than would have been otherwise possible. In this commentary, I briefly review the replication process and then consider its implications for VO, the decline effect, and the larger issue of using metascience to turn the lens of science onto itself.

The Unfolding of the VO Replication Study

In February 2013, I was asked if I would be willing to assist in developing and vetting the protocol for replicating Schooler & Engstler-Schooler (S&E-S) Experiment 1. I provided the replication team with the original stimulus materials and was given a detailed description of the protocol to review. The procedure was deceptively simple. Participants were to view a video of a simulated bank robbery, and subsequently they either would be asked to describe the appearance of the perpetrator or to engage in an unrelated activity (naming countries and capitals). Finally, all participants were to be given a line up and

Corresponding Author:
Jonathan W. Schooler, Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93105
E-mail: jonathan.schooler@psych.ucsb.edu

asked to indicate whether or not the perpetrator is present and, if so, to identify him. I made some minor modifications but unfortunately did not notice that the timing intervals were different from those used in S&E-S Experiment 1. In the original study, the verbalization manipulation was introduced 20 min after viewing the robbery video and immediately prior to test. However, in the protocol for the first study of the Registered Replication Report (henceforth "RRR1") verbalization occurred immediately after viewing the robbery video and 20 min before the test. In November 2013, I learned of the deviation in the procedure. Admittedly, the timing intervals included in RRR1 were similar to those used in S&E-S Experiment 4; in both, participants described the face immediately after viewing the video however S&E-S Experiment 4 included a 10-min rather than a 20-min interval between the description and the recognition test. Although the initial replication protocol resembled Experiment 4, I was wary of having it presented as a replication of S&E-S's VO paradigm as it did not entirely correspond to any of the originally conducted studies. Moreover, several lines of research conducted since the original S&E-S series (i.e., Finger & Pezdek, 1999; Meisner & Brigham, 2001) suggested that VO effects are maximized when verbalization occurs following a delay after encoding and immediately prior to test. It seemed like a bad idea to draw conclusions about the robustness of VO based on a protocol that did not entirely replicate any of the original studies and included timing intervals that subsequent studies suggested were not optimally suited to produce an effect. Fortunately, the replication team acknowledged my concerns and, prior to analyzing any of the results, initiated a second round of the replication process with the original parameters of S&E-S's Experiment 1.

The results of the two replication rounds strikingly demonstrated the value of including both timing variations. When verbalization was introduced immediately after viewing the robbery video and 20 min before the final test, there was only a slight VO effect (i.e., 4% difference in identification performance of participants in the verbalization condition relative to controls.) However, when the original parameters of S&E-S Experiment 1 were replicated in RRR2, the VO effect increased to 16%.

Implications for VO

The outcome of the replication effort clearly demonstrates that VO is a genuine phenomenon that can be quite substantial under the right conditions. Recently, Francis (2012) likened the VO effect to parapsychological findings, with the suggestion that it might simply be the product of publication bias. The magnitude and robustness of the VO effect across replication sites clearly rejects the speculation that VO is an artifact of selective publishing while also underscoring the potential theoretical and applied significance of this counterintuitive phenomenon. Theoretically, the existence of robust disruptive effects of describing a previously seen face suggests important boundary conditions on the efficacy of verbal rehearsal and the value of linguistic representations (Schooler, Fiore, & Brandimonte, 1997). Practically, these findings suggest that the standard forensic practice of soliciting eyewitness descriptions could, at least under some conditions, undermine effective identifications.

Although VO was quite substantial in RRR2 with timing parameters that mirrored those of S&E-S Experiment 1 (i.e., when verbalization occurred 20 min after viewing the face and immediately before test), it was markedly reduced in RRR1 when the order of verbalization and the delay were reversed. This difference between the two timing intervals conceptually replicates Finger and Pezdek (1999), who found that the negative effects of verbalization were attenuated when a delay was introduced between verbalization and test. Although the original VO studies featured no single experiment directly comparing different timing intervals, in contrast to the results of the replication study, several of the original VO studies observed substantial VO effects with varying intervals between verbalization and test (I will return to this point shortly in discussion of the decline effect).

The difference in the magnitude of the VO effect in the two RRR studies has several important implications for understanding VO. First, it highlights the susceptibility of VO to seemingly modest changes in experimental details. VO's susceptibility to relatively modest deviations in the paradigm's implementation may be one reason why it has been so challenging to replicate, as we can only speculate about whether there might exist other parameters of the VO paradigm (e.g., exposure duration, distractor similarity) that might similarly impact on its outcome. A second important issue that emerges from the difference between the VO effect observed with the two timing parameters is discerning the specific delay that moderates the effect. As the report notes, the two timing variations confounded the interval between encoding and verbalization with that between verbalization and test. From a theoretical standpoint, if VO depends on verbalization occurring immediately prior to test, this would support the notion that VO may result from the inappropriate transfer of cognitive processes (e.g., featural analysis) engaged in during verbalization to the recognition test (Macrae & Lewis, 2002; Schooler, 2002). From a forensic perspective, if the effects of verbalization increase with longer durations between encoding and verbalization then this would highlight the dangers of VO in eyewitness contexts (as witnesses are rarely interviewed immediately after the event). In contrast, if the delay between verbalization and test is critical, then the disruptive effects of VO may be less applicable in

forensic contexts as witnesses are rarely given line-ups immediately after describing a perpetrator.

Implications for the Decline Effect

Although the basic VO effect was robustly observed in RRR2 with the original timing parameters of S&E-S Experiment 1, there were a number of potential discrepancies between the findings of the replication effort and those of S&E-S. First, although RRR1 was not identical to S&E-S Experiment 4 (with a 20-min rather than a 10-min interval between verbalization and test) they were pretty similar. Nevertheless as the authors note "Whereas the original study showed a −22% difference between the verbal description condition and the control condition (verbal description – control), the meta-analytic effect across 31 larger scale replications was substantially smaller: −4.01% [95% confidence interval: −7.15% to −0.87%]" (p. 565). Second, the nature of the VO effect was somewhat different in the S&E-S experiments versus the replication studies. S&E-S Experiments 1 and 2, observed that verbalization resulted in comparable increases in errors involving incorrect identifications (false alarms) and "not present" judgments (misses). In contrast, the replication studies only observed an effect of verbalization on misses. Finally, although the magnitude of the VO effect observed in RRR2 (16%) was substantially larger than RRR1 (4%), it nevertheless was smaller than that observed in the original study (25%). In short, although the VO effect was unequivocally evidenced in the replication effort, it was neither as numerically large nor as broadly observed across timing variations and error types as it appeared in the original series of studies.

In my opinion, the numerical decrease in the VO effect size across timing parameters and error types observed in the replication effort is consistent with the general notion of a decline effect—the claim that reported effect sizes can diminish in magnitude over time.[1] Although decline effects of various sorts have been reported in a number of domains, including psychology (Clark, Moreland, & Gronlund, in press), biology (Jennions & Møller, 2002), medicine (Ioannidis, 2005; Kemp et al., 2010) and parapsychology (Bierman, 2001), their nature, source, and generality remain unclear. The present findings offer some possible hints about sources of the decline effect.

The most straightforward account of decline effects entails a combination of underpowered studies and regression to the mean. Researchers regularly conduct empirical investigations that are underpowered relative to the effect sizes that they report (Francis, 2012). This means that published findings must routinely be the beneficiaries of random variance favoring—or, at a minimum, being neutral toward—the reported result. However, when subsequent researchers attempt to replicate published findings, error variance will be randomly distributed on both sides of the true population mean, often leading to a reduced effect size relative to the originally reported result.

The results of this replication effort are in principle in keeping with the above account of the decline effect. The results of S&E-S Experiment 1 were larger than those of RRR2, but still within the meta-analytic confidence interval. Although the results of S&E-S Experiment 4 were outside of the meta-analytic confidence intervals of RRR1, regression to the mean could also account for this disparity as the *N* in the original study was relatively small. Regression to the mean could also explain why VO effects were observed with both false identifications and misses in the original study but only with misses in the replications. Again the *N* in the original study was small, so potentially they could have overestimated the situations in which VO actually occurs. Notably, the notion that the original studies overgeneralized the VO effect to false alarms when in fact it only applies to misses is a bit harder to square with the results of S&E-S Experiment 6, which found VO even when a not-present option was omitted as a response option, thereby exclusively limiting errors to false alarms. However, this study used a somewhat different paradigm (e.g., photographs rather than a video), so we must be cautious in drawing too strong conclusions.

Another account of the decline effect entails changes in procedures that are not originally recognized as being important. Evidence for the impact of underappreciated variables in mediating decline effects is also suggested in the present findings. The failure to accurately reproduce S&E-S's original timing parameters in RRR1 illustrates how easy it is to overlook slight changes in procedural details. The fact that these small timing permutations had such a large effect on the outcome of the results highlights just how significant such details can be. Although subtle changes in procedural details likely contribute to decline effects and certainly played a role in differences between in the magnitude of VO effects in RRR1 and RRR2, it is not clear that such considerations are sufficient to entirely account for decline effects. It is certainly possible that some as-yet unrecognized procedural changes can explain why S&E-S's Experiment 4 effect size was so much larger than that associated with the similar procedure of RRR1. However, it is unclear what those differences might be at present. Thus, the subtle-change-in-procedure account requires postulating the existence of unknown variables that may or may not exist.

In the end, although I recognize that decline effects can be accounted for by the mechanisms outlined above, I must (albeit reluctantly) acknowledge that I just cannot shake the sense that something else may be going on. Too many times have I gotten highly significant results

the first several times I tried a new paradigm and then found it increasingly challenging to get effects of comparable magnitude in later studies. It is notable in this regard that S&E-S Experiment 1 was the very first VO study that we ever conducted and yet somehow we managed to select precisely the right parameters for maximizing the VO effect. Variations of the procedure have produced effects but generally smaller than what we found the very first time. I have had similar experiences with a host of other paradigms including VO of insight problem solving (Schooler, Ohlson, & Brooks, 1993), analogical reasoning (Lane & Schooler, 2004), music memory (Houser, Fiore, & Schooler, 1997), and map memory (Fiore & Schooler, 2002). In each of these cases, we got effects more easily at first than we did subsequently. I certainly would not want to claim that this demonstrates that something as unconventional as beginner's luck is playing itself out in science. Nevertheless, I remain of the conviction that it is appropriate for scientists to entertain unconventional mechanisms even while maintaining a healthy skepticism about them. Perhaps, there are some parallels between VO effects and parapsychology after all, but they reflect genuine unappreciated mechanisms of nature (Schooler, 2011) and not simply the product of publication bias or other artifact. This is admittedly a provocative speculation, and one that many will dispute as unwarranted or worse. But I stand by the view, heralded by a minority of distinguished scientists (e.g., Bem, 2011; James 1902/2002; Penrose, 1989), that science can entertain and explore unconventional accounts without descending into the morass of superstition and irrationality.

Implications for Metascience

The outcome of the VO replication effort is in my opinion a genuine victory for the emerging field of *metascience*—the approach of turning the lens of science onto itself. By systematically enrolling researchers from all over the world to conduct a scientific investigation of a single scientific paradigm using a carefully regulated protocol, the authors were able to shed important light not only on the phenomena under investigation, but also on the strengths and pitfalls of the scientific process itself. Admittedly, the effort encountered a serious bump in the road by initially testing a protocol that did not constitute a full replication of the study it set out to emulate. However, even this glitch, once addressed, greatly enhanced how informative the enterprise ended up being. The present project provides an unprecedented example of the value of a carefully conducted, large-scale, multisite replication effort.

The success of this replication effort in fleshing out both the previously equivocal domain of VO and the endemic elements of science that may have contributed to VO's murkiness clearly highlight the value of this approach. However, this is just one of many metascientific tactics that is likely to bear great fruit. More generally, the practice of scrupulously logging research studies before they are conducted and reporting their results regardless of outcome (Mooneyham, Franklin, Mrazek, & Schooler, 2012; Nosek & Bar-Anan, 2012; Schooler, 2011) is certain to provide important insights into the nature of the scientific process. At present, only a fraction of all conducted research is ever reported, and the details of those studies that are published may be "finessed" to varying degrees (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). The thorough logging of procedures before findings are collected and the full reporting of data regardless of outcome afterwards will ultimately provide an unprecedented understanding of the scientific topics under investigation and will illuminate the process of science itself.

Finally, as researchers continue to investigate the process of replication, I hope that they will increasingly turn their attention to not only replicating studies that have already been reported, but also prospectively planning to replicate results that have yet to be discovered. Toward this end, research teams at UC Berkeley, Stanford, and the University of Virginia have joined with my lab (at UC Santa Barbara) in agreeing to examine the replicability of new findings that are uncovered while engaging in hypothesized "best practices" for maximizing the reliability of findings.[2] We will be carefully documenting and logging all aspects of our scientific protocols, using highly powered research designs, and then carefully replicating the protocols across universities regardless of initial replication success. This approach may elucidate how turning the lens of science onto itself can reveal practices for insuring robust effects and perhaps begin to reveal why seemingly large effects sometimes appear to decline in magnitude over time.

Acknowledgments

The writing of this commentary was supported by the Fetzer Franklin Trust. I thank Ben Mooneyham, Claire Zedilius, Daniel Simons, Brian Nosek, Bobbie Spellman, Alex Holcombe and John Protzko for comments on an earlier draft.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. Whether a decline effect is seen in the present data depends in part on how it is defined. As the authors note with respect to the discrepancies between RRR1 and S&E-S Experiment 4 "Although that original effect size estimate falls outside the

confidence intervals of our meta-analytic effect size for that study, it is unclear whether the effect actually declined in size or whether the original estimate was just an inaccurate estimate of the effect" (p. 570). In my view, a domain in which the original reported effect was an exaggerated estimate of the actual effect would still be an example of a decline effect, just one in which it was due to an initially over estimated effect size.

2. This multisite project is being supported by coordinated grants from the Fetzer Franklin Trust to each institution.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*, 556–578.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. doi:10.1037/a0021524

Bierman, D. J. (2001). On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective work of physics? In P. van Locke (Ed.), *The physical nature of consciousness* (pp. 269–292). New York City, NY: Benjamins.

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (in press). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin and Review*.

Finger, K., & Pezdek, K. (1999). The effect of the cognitive interview on face identification accuracy: Release from verbal overshadowing. *Journal of Applied Psychology*, *84*, 340–348.

Fiore, S. M., & Schooler, J. W. (2002). How did you get here from there: Verbal overshadowing of spatial mental models. *Applied Cognitive Psychology*, *16*, 897–910.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156.

Houser, T., Fiore, S. M., & Schooler, J. W. (1997). *Verbal overshadowing of music memory: What happens when you describe that tune?* Unpublished manuscript.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*, 218–228.

James, W. (2002). *The varieties of religious experience*. New York City, NY: Random House. Original work published 1902.

Jennions, M. D., & Møller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society, Series B: Biological Sciences*, *269*, 43–48.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Kemp, A. S., Schooler, N. R., Kalali, A. H., Alphs, L., Anand, R., Awad, G., . . . Vermeulen, A. (2010). What is causing the reduced drug-placebo difference in recent schizophrenia clinical trials and what can be done about it? *Schizophrenia Bulletin*, *36*, 504–509.

Lane, S. M., & Schooler, J.W. (2004). Skimming the surface: Verbal overshadowing of analogical retrieval. *Psychological Science*, *15*, 715–719. doi:10.1037/10590-011

Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker*, pp. 52–57.

Macrae, C. N., & Lewis, H. L. (2002). Do I know you? Processing orientation and face recognition. *Psychological Science*, *13*, 194–196.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*, 603–616.

Mooneyham, B. W., Franklin, M. S., Mrazek, M. D., & Schooler, J. W. (2012). Modernizing science: Comments on Nosek & Bar-Anan. *Psychological Inquiry*, *23*, 281–284.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Penrose, R. (1989). *The emperor's new mind*. Oxford, England: Oxford University Press.

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*, 989–997.

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature*, 470, 437.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71. doi:10.1016/B0-08-043076-7/01599-0

Schooler, J. W., Fiore, S. M., & Brandimonte, M. A. (1997). At a loss from words: Verbal overshadowing of perceptual memories. In D. L. Medin (Ed.), *The psychology of learning and motivation* (pp. 293–334) San Diego, CA: Academic Press.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*, 166–183.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.