6

Decline Effects
Types, Mechanisms, and Personal Reflections

John Protzko and Jonathan W. Schooler

It is tempting to believe that scientific findings provide an accurate account of enduring reality. The indisputable success of the scientific enterprise is testament to the significant degree to which initially reported findings can be replicated and built upon. Nevertheless, a substantial number of findings are less robust and less substantial than they initially appear (Chapters 1, 2, and 3). Some effects that were present have declined over time. Appreciation of the unreliability of scientific findings has led to what some have termed *the replication crisis*, as a variety of areas including biology (Begley & Ellis, 2012), psychology (Bakker, van Dijk, & Wicherts, 2012), and genetics (Siontis, Patsopoulos, & Ioannidis, 2010) have come to recognize – that a striking number of studies in their respective fields no longer replicate.

In this chapter, we consider four general types of declining effect sizes, each of which relates to the hypothetical true effect size of the finding in question at the time it was originally reported. *False positive decline effects* occur when there actually was no true effect when the research was conducted, initially reported positive findings were instead a statistical or methodological artifact. *Inflated decline effects* occur when a true effect did exist but the initially reported studies artificially inflated the estimate of its size. *Under-specified decline effects* occur when a true effect originally existed but its necessary conditions were under-specified, as a result subsequent studies failed to include those conditions and thereby observed smaller effects. Finally, *genuinely decreasing decline effects* occur when the true effect size was originally and accurately reported but, for some reason, the true effect genuinely declines in magnitude over time.

In documenting the various types of decline effects, we will depart from the standard approach of multi-author papers – of exclusively writing in a single collaborative voice. Certain sections of this chapter have been written by and

Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions, First Edition.
Edited by Scott O. Lilienfeld and Irwin D. Waldman.
© 2017 John Wiley & Sons, Inc. Published 2017 by John Wiley & Sons, Inc.

correspond to the opinions of only one author. Although we respect each other's opinions, the two authors have different perspectives on some central issues regarding the likelihood that unconventional mechanisms may play a role in science in general and the decline effect in particular. Protzko is skeptical of such claims, while Schooler believes they are worthy of consideration. Nevertheless, both share the view that decline effects have multiple sources, and that delineating those sources and the conditions under which they are likely to manifest is critical to making headway in this increasingly pressing topic.

Four Types of Decline Effects

Before expounding on the four distinct types of decline effects outlined in the preceding text, there are also a number of general mechanisms that may play a role in many of these cases. These most notably are artifactual sources that contribute to errors in effect size estimation, and include the following.

Underpowered studies

An important factor that can fuel declining effect sizes is the common tendency for studies to use underpowered designs. With smaller *N*'s, the probability greatly increases that a positive experimental result was inflated by error variance. A common difference between initial studies that show larger effects and subsequent studies that show smaller effects is the smaller sample size associated with the initial studies in a paradigm (Barto & Rillig, 2011; Button et al., 2013; Pereira, Horwitz, & Ioannidis, 2012). Since later studies use larger samples providing more conservative estimates, a decline effect emerges (Ioannidis, 2005; Ioannidis & Trikalinos, 2005; Ioannidis, Trikalinos et al., 2003).

Publication bias

Publication practices can create a decline effect through multiple routes (see Chapter 3). Publishing a novel finding can create a mini furor of research and commentary. Influential findings can sometimes create new paths of research to explore. During this time, fields generally become excited about a new finding and reject null results (Young, Ioannidis, & Al-Ubaydli, 2008). In effect, people do not want their new field to fall flat. This underreporting of failed replications can come from both the researchers and the editors. Researchers contribute to decline effects by not writing and submitting null findings (this also contributes to the file drawer problem). Even when researchers decide to submit null findings for publication, they take 1–2 years longer to write and submit completed results than they do for statistically significant results (Ioannidis, 1998). Editors contribute to decline effects by treating null

findings differently than they do statistically significant ones. After submission, it takes longer for editors to publish null findings than statistically significant results (Ioannidis, 1998). Statistically significant results reach the literature faster, while null trials, if they even make the literature, appear later.

Selective reporting

The final and arguably the most insidious artifactual source of decline effects is selective reporting. Given the incentive structure of academia and the high standards of select journals, a great temptation exists to cherry-pick dependent measures, covariates, and even conditions that produced sizable effects, while omitting those that weaken or complicate the story. Considerable evidence suggests that researchers routinely engage in selective reporting (John, Loewenstein, & Prelec, 2012), and that such practices may significantly contribute to replication difficulties (Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014). Moreover, given that the researcher who initially reports an effect will be identified with it, they may be less motivated to demonstrate the effect's magnitude and robustness, and thus incentivized to engage in selective reporting procedures. If initial researchers engage in a greater degree of selective reporting than replications, this too would fuel decline effects.

False positive decline effects

False positive decline effects occur when no true effect exists and subsequent scientific findings demonstrate that the initial finding was in error. This represents a regression to a true null mean. All of the mechanisms mentioned earlier are likely to contribute to false positive decline effects. In addition, some false positive decline effects may be due to errors in the initial procedures or analyses.

The Mozart effect provides one example of a false positive decline effect that seems likely to simply have been the victim of regression to the mean. In 1993, the first paper detailing the positive benefits of listening to Mozart was published (Rauscher, Shaw, & Ky, 1993). This first study compared students listening to Mozart's Sonata for Two Pianos in D major (KV 448) to students not listening to anything. Students who listened to KV 448 scored higher on a task of spatial ability. Replications of the Mozart effect with different conditions commenced. Some replications were successful (e.g., Rideout & Taylor, 1997), while others were not (e.g., Carstens, Huskins, & Hounshell, 1995; Steele, Bass, & Crook, 1999). Over time, the replication failures began to amass. It now seems there is no true effect of listening to Mozart on cognitive ability (Pietschnig, Voracek, & Formann, 2010). The initial findings appear to have been a statistical fluke.[1] The reason why later experiments were not finding an effect is presumably because there was never an effect in the first place.

Certain eyewitness identification procedures have undergone a similar decline, presumably due to regression to the mean. In a meta-analysis (Clark, Moreland, & Gronlund, 2014), the efficacy of four identification procedures that were originally found to produce no cost benefits to eyewitness identification (decreasing false identification while having no negative effect on correct identification) was tested. The four manipulations were (1) lineup instructions – comparing biased and unbiased lineups (Malpass & Devine, 1981); (2) lineup presentation – comparing sequential and simultaneous lineups (Lindsay & Wells, 1985); (3) lineup similarity – comparing more versus less similar filler members (Lindsay & Wells, 1980); and (4) filler selection method – comparing lineups with description-matched fillers to lineups with suspect-matched fillers (Wells, 1993). The results revealed that, in all four cases, the originally observed no-cost benefit of the manipulation attenuated over time. The true effect of such procedures incurs some increase in false identifications or some decrease in correct identifications. The early studies in eyewitness identification reform showed a *no-cost* effect to these interventions. However, in reality, it appears those procedures do incur a cost. The effect sizes declined over time, apparently because future replications were converging on the (true) null effect.

Facilitated communication is another example of a false positive decline. In this case, the failures to replicate were the products of improvements in methodology that revealed the flaws in the initial procedure. Facilitated communication was the methodology where a nonverbal patient – usually someone with dementia, autism, or in some degree of vegetative state – was paired with a facilitator who, using his or her training to respond to subtle movements of the patient, helped guide his or her hand over a keyboard that allowed the patient to communicate (Crossley & McDonald, 1980). In the world, however, there was little to no effect of facilitated communication (Jacobson, Mulick, & Schwartz, 1995). The facilitators were responding to what they saw, not what the patients saw. When the patient was asked to describe what they saw and were shown one picture but the facilitator saw a different picture (unknown to them), the patient would "respond" with what the facilitator saw (e.g., Bligh & Kupperman, 1993). Although the first results showed a large effect of facilitated communication, there was no true effect. Science converged on this null finding, with subsequent studies showing that the original effect was the result of an experimental artifact.

Another source of false positive decline effects is that the initially reported studies use inappropriate statistical methods. One example is that of Type D personality and heart disease. Someone who has a Type D personality often is negative and inhibited in social situations; these people are also more likely to die from heart disease (Denollet, Sys, & Brutsaert, 1995). This correlation between Type D personalities and death by heart disease, however, has been experiencing a decline effect (Coyne & de Voogd, 2012). The main reason proposed for this decline is changes in methodology. Initial studies finding an effect used median splits to determine who counted as socially inhibited and negative. Median splits are rarely if ever justified in scientific practice as they can increase the likelihood of Type I errors (DeCoster,

Iselin, & Gallucci, 2009). Later studies eschewing median splits were unable to find a relationship between a Type D personality and death by heart disease (Coyne & de Voogd, 2012), precisely because they were using more correct procedures.

Inflated decline effects

Although scientific artifacts can sometimes create false positive effects, many times a true effect exists but was artificially inflated. Inflated decline effects occur when a stable true effect exists but the effect is exaggerated due to the same sorts of factors (e.g., small *N*, selective reporting, publication bias) associated with false positive decline effects. The primary difference between false positive and inflated decline effects is whether the true effect exists.

There are a number of examples of what appear likely to be inflated decline effects stemming from artifactual factors such as publication bias, inadequate *N*, or regression to the mean. Notably, it has been suggested that the majority of all studies evince such patterns (Ioannidis, 2008; see Chapters 1 and 2). Some inflated decline effects appear to be due to either underpowered or poorly designed initial studies (see Chapter 4). In reviewing the sources for reasons why replications of medical studies tend to have smaller effect sizes than the original investigations, for example, studies associated with replications with diminished effect sizes were more likely to have smaller *N*'s and not include a randomized control group, relative to studies that were fully replicated (Ioannidis, 2005).

Changes in analyses can also create inflated decline effects. When secondary sexual characteristics are symmetrical in males (musculature, facial symmetry), they have an advantage in selecting mates (e.g., Møller & Thornhill, 1998). Over the years, there has been a decline effect; these characteristics are less likely to predict reproductive success across species (Simmons, Tomkins, Kotiaho, & Hunt, 1999). Newer studies on the role of symmetry in reproductive success use repeated methods that reduce measurement error (Björklund & Merila, 1997; Swaddle, Witter, & Cuthill, 1994) instead of single-exposure methods. These newer methods provide more accurate measures of the role of symmetry in reproductive success (Simmons et al., 1999), causing a decline as newer studies return smaller effect sizes.

Under-specified decline effects

So far, the decline effects we have reviewed involve situations in which the initial publications mischaracterized the magnitude of the true effect size. Under-specified decline effects, however, occur when the true effect is accurately characterized in magnitude but not with respect to the specifying conditions needed to observe it. In such cases, follow-up studies may fail to see comparably large effect sizes because they have inadequately reproduced the original conditions.

Some under-specified decline effects result from an under-specification of the population to which the effect generalizes. An excellent example of this type of decline effect occurs in online economic games. People give more money to a public pot under time constraints than if given time to think about how much to give (Rand, Green, & Nowak, 2012). This shows that people are naturally cooperative, and only when you give them time to think do they become greedy and selfish. The original researchers, however, could not replicate their own results (originally and subsequently done online). Exploring why this happened, they found that an online subject participates in more economic games in one week than real-life laboratory subjects complete in their entire careers (Rand et al., 2013). Some online subjects even report participating in thousands of economic games. Using participants with more experience in economic games makes the time constraint effect on giving disappear; when the researchers used only subjects who are new to economic games, they replicated their original finding (Rand et al., 2013). In this, the true effect remains stable, but the researcher did not know and hence did not report the population specifications (i.e., minimal experience with economic games) necessary to observe the effect.

Decline effects due to under-appreciation of the necessary population specifications occur in other fields as well. The carbon–nutrient balance (CNB) theory in ecology predicts that plants alter their nutrient concentrations in response to being eaten (Karban & Myers, 1989). The evidence for the CNB theory exhibited a decline (Nykänen & Koricheva, 2004), appearing to no longer be a true effect. What was happening with the CNB theory was a change in the types of plants studied. The most common plant first studied was the Scots pine (often used as Christmas trees); as the research progressed, new plant species were studied, leading to the appearance of a decline effect (Leimu & Koricheva, 2004). The CNB theory is robust when studying Scots pine, but does not generalize to all plants. Again, changes in the subjects created an under-specified decline effect.

The medical field also experiences under-specified decline effects due to changes in the specifications of the populations from which the samples were drawn. Over time, the effectiveness of the drug Timolol to treat glaucoma decreased (Gehr, Weiss, & Porzsolt, 2006). This same study also showed declining effects of the drug Pravastatin for lowering lipids. On inspection, the decline likely occurred because later research on Timolol and Pravastatin included patients who were not as advanced in their respective diseases as the earlier studies (Gehr et al., 2006). The first studies used patients with advanced glaucoma and heart disease, for which the drugs worked. Later studies used less advanced patients, for which there was less room for improvement. The change in sample characteristics apparently led to the decline effects in these studies.

Genuinely decreasing decline effects

As discussed, a variety of factors can create the appearance of a lessening true effect size over time. Indeed, we assume that any variation in effect sizes over time is non-systematic. Effect sizes bounce around because they are randomly drawn from an

effect size distribution. This distribution has mean θ (what a meta-analysis seeks to uncover), but such a mean is generally assumed stable in the world. There have been a few studies, however, that suggest that their local θ may not be stable, and that the variation is systematic and declines over time. Several accounts have been offered for declines in which the true effect size appears to genuinely decrease. The most straightforward account is changes to the population.

One interesting example of a genuinely decreasing decline effect that is likely due to changes to the population comes from work on a parasite's ability to alter the behavior of its host to increase transmission. Certain types of tapeworms, for example, infect brine shrimp, turning them bright red and making them swim nearer the surface; all this so birds (such as flamingos) will be more likely to eat the shrimp. and the tapeworm can infect its target host (Sánchez, Georgiev, & Green, 2007). In a meta-analysis of studies supporting this host-manipulation paradigm, over time, infected shrimp exhibited less behavioral changes over time (Poulin, 2000). Over the years, the sample sizes have not changed, and all of the studies were statistically significant. Therefore, this decline is unlikely to be driven by changing sample sizes, changing sample characteristics, or biased publication.

A variety of genuinely decreasing decline effects appear to stem from cultural developments. White students' tendency to attribute negative traits to *all* African-Americans was high in the 1930s (Katz & Braly, 1933), declined in the 1950s (Gilbert, 1951), and continued to decline in the 1960s (Crowne & Marlowe, 1964; Karlins, Coffman, & Walters, 1969). In the 1970s, it was discovered that some of this fading was due to increased social desirability of responses, but there was still a genuine and continuing decline in people's endorsing prejudicial statements (Sigall & Page, 1971). Following the same procedures, a continuing decline was apparent in the 1980s as well (Dovidio & Gaertner, 1986). Changing social conditions has taken what was initially a large effect and made it decline over time.

This does not mean that prejudice itself had been declining. Awareness of such stereotypes appears to have remained stable, while the *endorsing* of such stereotypes has been in decline (Devine, Monteith, Zuwerink, & Elliot, 1991). In addition, some stereotypes may be increasing over time but the content has become decidedly more favorable to past stereotypes (Madon et al., 2001). These results are contingent on how the stereotypes are asked and recorded by the participants as well, lending more complexity to the issue than previously perceived (Plant, Devine, & Brazy, 2003). It appears, however, that explicit endorsement of negative stereotypes of African-Americans has reduced in America since the 1930s (see Chapter 10).

Changes in stereotypes can also have implications for other fields of research, creating further genuinely decreasing decline effects. One such example is findings on stereotype threat for girls in math. A long-standing stereotype in many Western countries is that males are better than females in math (Fennema & Sherman, 1977; Nosek et al., 2009). This led to math being a threatening subject for girls. When taking a math test, girls perform worse if reminded, explicitly or implicitly, about the stereotype (e.g., Spencer, Steele, & Quinn, 1999). This effect, however, is going away. No longer are girls even aware of the stereotype that they are supposed to be worse

than boys at math (Plante, Theoret, & Favreau, 2009). Girls are outperforming boys in every facet of school, including math (Cole, 1997). This has led to boys now being stereotyped as bad in school; reminding elementary school children of this stereotype has been shown to cause a decrease in boys', not girls', math performance (Hartley & Sutton, 2013).

Indeterminate and non-conventional decline effects

Although in principle all decline effects can be categorized into one or more of the above four classes, in practice such classifications may be difficult without knowing the precise source of the decline. Indeed, in some cases, researchers openly acknowledge some mystification over the cause of the diminishment of the effect in question. For example, between 1993 and 2006, the effect of antipsychotics steadily declined among randomized-placebo-controlled designs (Kemp et al., 2010; Chapter 13). An investigation into possible reasons was undertaken, and it was proposed that the reason could be such factors as repeat subjects in multiple trials, participant characteristics, site characteristics, and trial designs. None of these solutions was immediately accepted, as there was not a systematic observation of those forces across studies.

The effectiveness of cognitive behavioral therapy (CBT) treatments have also been steadily declining since they were first introduced (Johnsen & Friborg, 2015). A number of possible sources for declining CBT effects have been conjectured, including laxer adherence to the specific therapy regimen and reduced patient expectations. However, as the authors note, these factors might reasonably have been expected to be counteracted by improvements in therapy delivery.

The impact of transcranial direct current stimulation (tDCS) on neuromodulation of brain activity has also undergone a gradual decline whose source has proven difficult to identify. Horvath, Forte, and Carter (2015) reviewed a variety of possible reasons for this decline, including possible changes in the duration of stimulation, the use of double-blind procedures, or the reliance on neuronavigation. None of these technological factors, however, accounted for the diminishing effect of tDCS over the last 14 years; as in the case of CBT, the authors noted that methodological advances could reasonably have been expected to enhance the observation of reliable effects.

Given the frequent lack of definitive evidence for the source of decline effects, some (including the second author, Schooler, 2011) have speculated about the possible involvement of mechanisms that are more non-conventional (see also Bierman, 2001). In a commentary in the journal *Nature*, Schooler (2011) mentioned the assorted conventional sources of decline effects detailed here, but also conjectured about the possibility of something more remarkable, noting:

> Less likely, but not inconceivable, is an effect stemming from some unconventional process. Perhaps, just as the act of observation has been suggested to affect quantum measurements, scientific observation could subtly change some scientific effects. (p. 437)

According to this view, even when all other variables are held constant, the mere repeated observation of an effect may be sufficient to induce a decline. Although the two authors of this chapter disagree about the likelihood that unconventional mechanisms of this sort may affect the decline effect, they concur that these represent a testable conjecture. To discover if a decline effect represents a genuine diminishment in the effect due to non-conventional mechanisms such as observation, researchers must make multiple observations over time that: (a) fully replicate the procedure; (b) maintain the same sample sizes; (c) sample from the same populations; and (d) use the same analytical methods. We agree that decline effects found under these conditions would constitute evidence that some non-conventional mechanisms, such as the act of observation, contributes to the phenomena, but we disagree about the likelihood that decline effects would be found under these highly controlled circumstances.

Separate Reflections on Unconventional Sources of Decline Effects by Schooler and Protzko

Reflections by Schooler[2]

Although hopeful that conventional accounts[3] may be sufficient to explain all decline effects, several considerations lead me to keep the door open to more unconventional accounts. Many readers are likely to recoil at this suggestion. Why would a reputable scientist speculate about mechanisms that challenge our current understanding of science, when aware of conventional mechanisms that could in principle account for all of the findings? I think that this is an understandable reaction, and indeed (as evidenced by the nature and co-authorship of this chapter) I fully respect those who conclude that my intuitions on this matter are off base. However, I believe that science flourishes when infused with alternative testable conjectures. Although my speculations may challenge current scientific tenets, they are falsifiable, and thus open to rational scientific evaluation. Indeed, efforts to explore these hypotheses could well refine the rigor of the scientific method, even if they do not reveal any of the anomalies that I entertain as possibly involved in decline effects. Even if I am entirely wrong in my conjectures, efforts to falsify them are likely to be useful. Furthermore, if there were something to these (albeit unlikely) conjectures, they would be of historical significance.

Before engaging in the specifics of my concerns, let me address one additional guiding theme of science that could reasonably be invoked at this juncture: the principle of parsimony (otherwise known as Ockham's razor). Generally, when adjudicating between alternative accounts, the explanation with the fewest assumptions is the most likely to be accurate. Given the efficacy of this principle, why entertain accounts that call upon unknown mechanisms, when simpler explanations are available? In this context, it is helpful to consider the words of Einstein (1934), who observed:

> It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience. (p. 165)

Although history routinely illustrates the value of parsimony, on occasion, long-held explanatory systems have proven inadequate in explaining seemingly small anomalies. For example, at the turn of the century, Newtonian physics seemingly explained virtually all known physical phenomena, with the exception of the orbit of mercury. Clearly, parsimony favored the view that this small anomaly could be accommodated within the Newtonian framework, and initially it was assumed that it could be (e.g., an additional unseen moon). However, in the end, Mercury's orbit (along with several other obscure anomalies) proved to be a telltale sign of the need for a whole new realm of explanatory mechanism: relativity theory (Einstein, 1920/2001). There are, of course, many other examples in science, where challenging findings were ultimately accounted for without major scientific re-conceptualization. However, the lessons of history illustrate the value of remaining open to the possibility that current scientific anomalies may require explanatory shifts of a magnitude rivaling those signaled by the slight deviations in Mercury's orbit.

My concern with standard accounts of decline effects is that there are several nagging "data of experience," both as they appear in the literature and as I have witnessed in my own research, that I am not entirely persuaded can be accounted for within standard frameworks. First, although conventional mechanisms can in principle account for all decline effects, in many cases, the demonstration of the causal relationship has yet to be established, and, in some cases, researchers remain largely in the dark as to the source. Second, I am struck by the fact that a large proportion of decline effects (virtually all reviewed in this chapter) exhibit a gradual decrease in effect size over time. Many standard mechanisms (e.g., regression to the mean, selective reporting) can explain why initial results would be inflated. However, they are less straightforward in explaining why effects continue to decline over time, often in a quite linear manner.

Admittedly, there are several conventional mechanisms that are likely contribute to at least some gradual decline effects, including population change, systematically investigating new populations for which the effect is increasingly unlikely to be observed, refinements in methodology, and the use of increasingly larger N in later experiments. While such mechanisms are likely involved in some gradual decline effects, at present, no study has demonstrated that they are sufficient to account for all such declines. Indeed, studies that have attempted to isolate individual variables have shown decline effects even when the critical variable was factored out. For example, in one of the most complete decline effect meta-analyses (including 44 peer-reviewed meta-analyses in ecological and evolutionary biology), Jennions and Møller (2002) found a gradual linear decline effect even when controlling for the larger *N* of later studies.

Moreover, there are a host of mechanisms that should contribute to the observation of increasingly *larger effect sizes*. Given the premium for positive results, over time, researchers might reasonably be expected to refine their paradigms in order to identify populations, methodologies, and necessary sample sizes that would maximize the likelihood of robust effects. In short, while extant conventional mechanisms may account for the consistent gradual decline effects that are routinely observed

across domains, the current state of evidence has yet to document this claim. From my vantage, the ubiquitous observation of unexplained gradual decline effects across disparate domains represents an unexpected anomaly that, like the anomalous orbit of mercury, may not be as easily accommodated within the extant scientific framework as it first appears.

My hunch regarding the possible involvement of unconventional mechanisms is further fueled by research in my lab, where I have repeatedly observed initially large effects wane, both in magnitude and in the various contexts in which they are observed. For example, in 1990, Tonya Engstler-Schooler and I found that participants who described the appearance of the perpetrator they had seen in an earlier videotaped depiction of a bank robbery exhibited recognition rates that were 25% less accurate than those who did not describe the perpetrator. Five variations of this experiment produced comparably large "verbal overshadowing" effects (Schooler & Engstler-Schooler, 1990). However, subsequent verbal overshadowing studies were less consistently successful. Some did not work at all (and were put in the file drawer); others produced significant effects that were substantially smaller than the original findings (Ryan & Schooler, 1998). A meta-analysis of studies using the verbal overshadowing paradigm (Meissner & Brigham, 2001) concluded that the effect was real, but markedly smaller than what we had routinely found in our early studies. Moreover, although we found verbal overshadowing effects in other domains including taste (Melcher & Schooler, 1996), music (Houser, Fiore, & Schooler, 1997), voices (Schooler, Fiore, & Brandimonte, 1997), insight problem-solving (Schooler, Ohlsson, & Brooks, 1993), artificial grammar (Fallshore & Schooler, 1993), and analogical retrieval (Lane & Schooler, 2004), later unpublished findings were, in all of these cases, smaller and less robust than the initial ones.

Recently a large-scale replication project including over 30 labs sought to replicate the original verbal overshadowing effect (Alogna et al., 2014). Although it produced highly significant findings, the overall magnitude of the effect was smaller than that observed in the original studies. Moreover, variations in the timing parameters that had no impact on performance in the original study led to a virtual disappearance of the effect in the replication studies (for a discussion, see Schooler, 2014b). I recognize that the apparent reduction in the verbal overshadowing effect in the replication studies relative to the original studies could have been due to regression to the mean, the smaller *N* in the original experiments, and/or differences in the precise manner in which the experiments were conducted. I also appreciate that our original ability to find verbal overshadowing with a host of timing parameters may have represented false positive effects. Nevertheless, I cannot escape the sense that it was somehow originally easier to get verbal overshadowing effects than it is today.

Importantly, decline effects are not the only "datum of experience" that may challenge conventional accounts of the role of the observer in science. Although effects of experimenter expectations on the outcome of studies have been observed for years (Rosenthal, 2005), we still do not fully understand the mechanisms

underpinning them. In commenting on the possible role of unconventional mechanisms in experimenter expectancy effects, Robert Rosenthal (the pioneer of this field) observed:

> Gordon Allport also believed that interpersonal expectancy effects might well be mediated parapsychologically. As of today, I have no evidence to support that position, nor do I have evidence to support the position that parapsychological phenomena are not involved in the mediation of interpersonal expectancy effects. Over the years, my students and I have found a number of potential mediating variables, but we are a long way from explaining all of the mechanisms that serve to mediate the operation of interpersonal expectancy effects. (Robert Rosenthal, personal communication, 11/14/11)

Inadequately understood observation effects are also famously found in physics, where the manner in which energy is measured appears to influence the form (particle or wave) in which it manifests. Although physicists have long been aware of the seeming impact of observation at the quantum level, there remains no consensus regarding its source (Schlosshauer, Kofler, & Zeilinger, 2013). Indeed, science does not even have a clear understanding of what it means to be an observer. It seems reasonable to argue that observation requires a conscious observer, as the outcome of any measuring device remains unknown until some conscious entity takes note of it. Yet, we remain largely in the dark regarding what consciousness is or how it relates to the physical universe (Chalmers, 2002; Schooler, 2015). Although many believe that the so-called "hard problem of consciousness" will eventually be solved by conventional mechanisms, few claim to have solved the problem, or to even be able to conjecture about what a solution might look like.

Given the host of unknowns surrounding the decline effect in particular and the process of observation more generally, it seems appropriate to maintain humility about how these vexing questions will be answered. To be sure, conventional mechanisms may be adequate to account for all current and future decline effects. Nevertheless, it remains possible that some mechanisms outside of our standard explanatory system will be involved.[4] The last century has been replete with a number of conceptual revolutions in understanding how the universe operates, most of which were first intimated by small anomalies. Given recent history, there seems every reason to think that there may be additional major paradigm shifts out there, particularly when it comes to the role of the observer in physical reality. The two largest scientific upheavals of the past century (relativity theory and quantum mechanics) both critically entailed gaining new understandings of the role of the observer. Indeed, the current inability of science to adequately situate the observer in extant models of physical reality is itself sufficient to suggest that further major scientific revolutions may be under foot (Schooler, 2015). The decline effect, with its potential relevance to the process of observation, resides within a particularly ill-understood scientific realm that seems especially ripe for major reconceptualization.

Fortunately, this is a debate that can be resolved by science. If observation itself contributes to decline effects, then they should be impacted by the manner in which

scientific findings are recorded by a conscious observer. Similarly, if genuine effects diminish as a function of repeated observation, then seemingly false positive decline effects may actually correspond to real phenomena that have undergone genuinely decreasing decline effects with respect to the boundary conditions under which they can be observed. In other words, initially promising empirical findings that seem to have diminished to the point that they no longer appear genuine, may (at least sometimes) have been prematurely dismissed. Rather than being false positives, they may, like verbal overshadowing, correspond to real effects that are smaller and/or more circumscribed than they originally appeared. The Mozart effect, the benefits of sequential vs. simultaneous lineups, and the impact of personality on heart disease might actually be true effects whose boundary conditions have become more delimited, and thus easier to fall outside of. If this radical speculation is right, then systematically investigating alternative boundary conditions for seemingly false positive effects may find the "sweet spot" – that is, the particular combination of parameters (like those discovered in the large-scale verbal overshadowing replication) where the effect still resides. These may be far-fetched predictions, but they are falsifiable, and, thus, particularly given their potentially monumental implications, an appropriate domain for further scientific inquiry.

Reflections by Protzko[5]

Based on the literature, the effectiveness of a research outcome can appear to decline over time. We have outlined what we believe are the scientific causes of such declines, including regression to the mean, changing populations, and changing analytic strategies. The question that remains is what to make of genuine declines in a true effect despite these changing procedures (genuinely decreasing decline effects). Some of these decline effects are straightforward: with girls outperforming boys in school, the stereotype that girls are worse at math than boys *should* go away, along with effects that are dependent on such a stereotype (such as gender stereotype threat). What I believe may be happening with other decline effects that have no such ready answer is a combination of confirmation bias and the incentive structure of academic science.

Assume one were able to view a mega-analysis (meta-analysis of meta-analyses) of every study ever done, organized by the specific procedure/experimental paradigm. Even controlling for the causes we outline of potential decline effects (e.g., changes in populations, changes in analytic strategy, changes in sample size), there would still be *random fluctuation* of effect sizes over time. Some effects would decline over time (tapeworms affecting brine shrimp coloring and behavior; Poulin, 2000). Some effects would incline over time (larger effect of exposure to mass media on girls' ideal weight; Grabe, Ward, & Hyde, 2008). Some would behave in truly strange ways over time (heritability of intelligence in Norway, see Figure 6.1; Sundet, Tambs, Magnus, & Berg, 1988). Most, however, would remain relatively stable given an absence of the effects discussed previously (e.g., effectiveness of creativity training; Scott, Leritz, & Mumford, 2004).

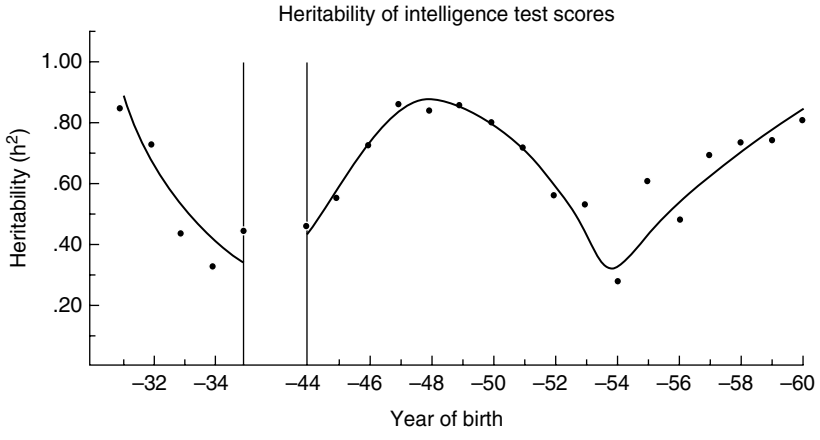

Figure 6.1 Changes in the heritability of intelligence for yearly cohorts of almost every male in Norway when they are 18 years old. From Sundet et al., 1988.

Under this representation, there would be a number of effects that *exhibit* a decline effect, but, in the global scheme, most of this change would be random – a Type I error. So why do we focus on the declining effects and not the inclining ones, the strange ones, or the unchanging ones?

The incentive structure of academic science is one where a researcher is most rewarded for building a career on the discovery of a new effect. This effect becomes theirs, for example, *Schooler's* verbal overshadowing effect. Replication does not make a career. The discovery of an effect makes a career. Incline effects are not discussed because they only go on to reinforce the existence of the effect. Stable effects are not discussed because they are uninteresting. Decline effects, however, have all the intrigue of a murder mystery. Why the decline? Was there some nefarious behavior on the part of the experimenter? Academic fraud? Was it always a Type I error? What does this mean for the reputation or standing of the discoverer? These ideas capture us and lead us to give substantial interpretation to what may be just a Type 1 error of our mega-analysis. Therefore, we look for decline effects, ignoring unchanging or inclining effects. This is a form of confirmation bias. There has been no frenzy over the "replication marvel" when we find an increase in the effect over time.

Where Schooler and I agree is that, regardless of the cause of a genuine decline effect, be it as boring as my mostly Type I error explanation or as fantastic as Schooler's unconventional effects, the question is a scientific one. It demands a scientific answer. This has lead both of us into the field of meta-science.

Meta-Science and the Empirical Unpacking of the Decline Effect

Although we disagree regarding the likelihood that genuinely decreasing decline effects are common and/or mediated by unconventional mechanisms, we concur that the best way to move forward in understanding decline effects is through science.

Although increasing awareness of the challenges of scientific replication has been characterized as a "crisis" in science, we see it as heralding an exciting new era of "meta-science" (Schooler, 2014a, 2014b) in which the lens of science is turned squarely on itself. Numerous scientific endeavors have recently arisen that are likely to offer deep insights into the extent and source of decline effects. Large-scale replication efforts (Simons, Holcombe, & Spellman, 2014) are beginning to determine the extent to which extant scientific findings are robust, offering clues as to the types of findings that are more versus less likely to replicate. New statistical approaches (Simonsohn et al., 2014) are helping to identify the characteristics of studies that may have undergone the type of partial reporting practices that are likely to contribute to decline effects. The open-source pre-registering of experimental paradigms before they are conducted, and logging of outcomes afterward, is quickly turning from a pipedream (Schooler, 2011) to a reality that is supported by both a major open science platform (http://centerforopenscience.org) and top-tier journals (e.g., http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/badges) (see Chapters 1 and 5).

A number of important directions will need to be explored in order to gain a better handle on decline effects. Above all, more comprehensive meta-analyses across scientific fields would be invaluable for understanding the proportion of scientific effects that decline, incline, or remain steady, and the factors that contribute to these differences. Although we have focused on decline effects in this article, many studies show no systematic trends of the effect sizes over time (Capon, Farley, & Hoenig, 1990; Gehr et al., 2006; Grabeani, Rizos, & Ioannidis, 2007; Kayande & Bhargava, 1994, studies 3 & 4; Scott et al., 2004; Tellis & Wernerfelt, 1987; Tu, Tugnait, & Clerehugh, 2008). Incline effects have also been observed in a number of domains. Some incline effects are straightforward. Certain medical procedures are becoming more effective (e.g., the effects of chemotherapy on non-small-cell lung cancer; Ioannidis, Polycarpou et al., 2003), and certain social sensitivities are becoming more pronounced (e.g., women's responses to mass media that the ideal body shape of a woman is thin; Grabe et al., 2008). In some cases, however, it is hard to understand why incline effects have been observed. Before 1988, the heritability of sexual ornamentation (physical traits like a peacock's feathers that distinguish one member over the other males) was 0.37; however, from 1988 to 1996, the heritability rose to 0.67 (Alatalo, Mappes, & Elgar, 1997). Clearly, understanding the implications and magnitude of decline effects requires more field-wide analyses to determine the degree to which decline effects represent a disproportionately large tendency of scientific results over time.

A second crucial requirement for a deeper understanding of decline effects is the adoption of protocols that lead to greater transparency in science (Chapter 5). At present, many scientific studies (no one knows what proportion) are never reported, and those studies that are reported often represent only a portion of the measures, conditions, and/or analyses that were used (Chapter 3). It is unclear exactly how this widespread selective reporting affects the pattern of outcomes over time; it may contribute both to the occurrence of decline effects and to the obfuscation of their causes (Schooler, 2011). One important remedy to the current lack of

transparency in science would be the adoption of pre-registration and open data sharing of all studies, both published and unpublished. Greater access to the process and products of scientific research would illuminate both the scientific practices that affect the replicability of findings and the overall frequency with which initially discovered findings decline over time.

Finally, replication studies need to be devised that systematically investigate specific hypotheses regarding the factors that may contribute to decline effects. Recently, we initiated a multi-site prospective replication study to investigate how newly discovered findings fare upon repeated replication. Research teams at UC Berkeley, Stanford, and the University of Virginia have joined with our lab (at UC Santa Barbara) to examine the replicability of new findings that are uncovered while engaging in hypothesized "best practices" for maximizing the reliability of findings. This project (supported by the Fetzer Franklin Fund) is carefully documenting all aspects of newly developed scientific studies, using highly powered research designs, and then repeating the studies at the various universities. Such prospective replication experiments may illuminate the factors that govern the replicability of scientific findings, including: researchers' investment in the hypothesis, the number of times a protocol is repeated, and the manner in which methodologies and outcomes are communicated. This project can even begin to test non-conventional accounts of the decline effect, as every study will be run in two identical successive blocks. By analyzing each block separately and varying whether the temporally first or second block is analyzed first, we can begin to assess whether there is any impact on outcome of the time at which a study is run (or even less likely) when it is analyzed.

Although much remains to be learned about the factors that underpin the replicability of scientific findings, it is an exciting prospect that science can be used to address its own limitations. Of course, efforts to understand declining effects are not without risks. It is easy to perceive replication efforts as a personal attack on one's scientific credibility. Although recent advances may encourage researchers to avoid practices (e.g., cherry picking, *p*-hacking, using underpowered designs) that are associated with unreliable findings, we must avoid perceptions that replication efforts are for weeding out sloppy scientists. It would also be well advised to include, in replication efforts, additional measures or manipulations that can advance the programs they are investigating.[6] Although pre-registering procedures and logging results regardless of outcome are likely to provide deep insights into the sources of replication difficulties, care should be taken to ensure that such efforts are not stifling. Creative scientific advances can depend on researchers' willingness to engage in high-risk studies and to explore analytical strategies that they had not thought of at the time the study was implemented. Consideration should be given to how to best balance the needs of fostering the transparency of science with that of protecting scientists' capacity for creative and flexible investigation. As with all major scientific innovations, some are likely to question the merit of turning science on itself; however, with sufficient thought and rigor, it seems inevitable that meta-science will make inroads in explaining when findings replicate and when they decline.

Acknowledgments

The writing of this chapter was assisted by the Fetzer Franklin Fund, which provided support to both authors and sponsored a meeting on the decline effect in 2013 at UC Santa Barbara that helped to further many of the issues discussed here. We would also like to thank Drew Bailey for some insights that were incorporated.

Endnotes

1 The authors differ in their respective certainty that the original findings associated with this and several of the other studies listed in this section were merely false positive effects. Protzko is confident that these initial effects were simply Type 1 errors. Although Schooler concurs that this is a reasonable account, he remains open to the speculation that the effects were actually present initially but for some reason became harder to find over time. (See Schooler's discussion of this speculation on page 15.)

2 Citations to material included in the section under the header "Reflections by Schooler" should be in this format: Protzko and Schooler (2015; Schooler's personal reflections on the decline effect). The reference list entry should be in this format: Protzko, J. and Schooler, J. W. (2015). "Decline effects: types, mechanisms, and personal reflections." In Scott O. Lilienfeld and Irwin D. Waldman (Eds.), *Psychological Science Under Scrutiny* (pp. 87–109). Chichester, UK: Wiley.

3 Let me mention one additional (at least semi-conventional) mechanism that I think may play an important role in some underspecified decline effect in psychology: namely, whether the experimental conditions encourage an intuitive or analytic mode of processing (Epstein, Lipson, Holstein, & Huh, 1992). In attempting to resolve why terror management effects (e.g., Greenberg et al., 1990) often failed to replicate, Simon et al. (1997) varied whether the experimenter was formal or informal in appearance. They found that encouraging participants to think about death only triggered worldview defenses when the experimenter was informal. Their account of this finding was that informal experimenters induce a more intuitive mode of processing (Epstein et al., 1992) that enables unconscious defense mechanisms, whereas more formal experimenters lead to analytic processing that minimizes such unconscious processes. In a similar manner, it seems plausible that at least some psychological effects (e.g., unconscious goal priming) that have been characterized as false positives (e.g., Pashler, Coburn, & Harris, 2012; Pashler, Rohrer, & Harris 2013) may instead reflect under-specified decline effects resulting from the original studies' critical reliance on experimental contexts that encourage an intuitive mode amenable to the effects of unconscious processing.

4 It is possible to recognize the existence of non-conventional mechanisms without being able to adequately explain them. Indeed, this is very much the current situation with the effects of observation in quantum mechanics where physicists recognize that they challenge current conventional accounts but have yet to adequately explain them (Schlosshauer et al., 2013). If evidence arises to support the possibility of non-conventional accounts of decline effects, serious thought will need to be devoted to what might be going on. One albeit far-fetched suggestion is that something akin to beginner's luck may be present in scientific inquiries (Schooler, 2014b). When researchers investigate a domain for which a real effect is possible, some type of ubiquitous affordance of nature may make

that effect easier to spot initially than it is subsequently. An analogy for my admittedly far-fetched conjecture may be useful. Imagine that we were to point a very powerful telescope toward a distant object. The telescope is initially unlikely to be perfectly focused on the distant object. As a consequence, the image of the object will occlude a larger visual angle (i.e., appear bigger and fuzzier) than it would if the telescope were perfectly focused. As the telescope is brought into focus, the object will become more clearly demarcated but it will also become smaller (as the surrounding fuzziness is diminished). If the telescope were not aimed directly at the object but rather off a bit to one side, it is possible that, in the process of focusing the telescope, the object could disappear from view entirely. I conjecture that something similar may be going on with the decline effect. When researchers discover a new region of interest in the information space that constitutes reality, our metaphorical observational telescopes are necessarily out of focus, making the region appear larger and blurrier. As we conduct additional investigations we bring phenomena into better focus, but this means they no longer fully appear in all the regions that they once did.

5 Citations to material included in the section under the header "Reflections by Protzko": Protzko and Schooler (2015); Protzko's personal reflections on the decline effect). The reference list entry should be in this format: Protzko, J. and Schooler, J. W. (2015). "Decline effects: types, mechanisms, and personal reflections." In Scott O. Lilienfeld and Irwin D. Waldman (Eds.), *Psychological Science Under Scrutiny* (pp. 87–109). Chichester, UK: Wiley.

6 It is notable that one of the most important discoveries to emerge from the verbal overshadowing replication effort, namely the impact of temporal parameters, resulted from an error in the initial protocol. Building a conceptually interesting variable into replication efforts would enable other projects to similarly advance the understanding of the paradigm in question. Another useful approach would be if each replication team included some additional variable or measure in their individual replication project. Such embellishment of replication studies could enable them not only to determine whether the phenomenon under investigation is genuine, but also to further its more general understanding.

References

Alatalo, R. V., Mappes, J., & Elgar, M. A. (1997). Heritabilities and paradigm shifts. *Nature*, *385*, 402–403.

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.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Barto, E. K., & Rillig, M. C. (2012). Dissemination biases in ecology: Effect sizes matter more than quality. *Oikos*, *121*(2), 228–235.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Bierman, D. J. (2001). On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective world of physics. *The Physical Nature of Consciousness*, 269–292.

Bligh, S., & Kupperman, P. (1993). Evaluation procedure for determining the source of the communication in facilitated communication accepted in a court case. *Journal of Autism and Developmental Disorders*, *23*, 553–557.

Björklund, M., & Merilä, J. (1997, January). Why some measures of fluctuating asymmetry are so sensitive to measurement error. In *Annales Zoologici Fennici* (Vol. 34, No. 2, pp. 133–137). Helsinki: Suomen Biologian Seura Vanamo, 1964.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Capon, N., Farley, J. U., & Hoenig, S. (1990). Determinants of financial performance: A meta-analysis. *Management Science*, *36*(10), 1143–1159.

Carstens, C. B., Huskins, E., & Hounshell, G. W. (1995). Listening to Mozart may not enhance performance on the revised Minnesota paper form board test. *Psychological Reports*, *77*(1), 111–114.

Chalmers, D. J. (2002). Consciousness and its place in nature. In D. Chalmers (Ed.), *Philosophy of mind: Classical and Contemporary*. Oxford, UK: Oxford University Press.

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review*, *21*(2), 251–267.

Cole, N. S. (1997). *The ETS gender study: How males and females perform in educational settings*. Princeton, NJ: Educational Testing Service.

Coyne, J. C., & de Voogd, J. N. (2012). Are we witnessing the decline effect in the Type D personality literature? What can be learned? *Journal of Psychosomatic Research*, *73*(6), 401–407.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive*. New York: Wiley.

Crossley, R., & McDonald, A. (1980). *Annie's coming out*. Middlesex, England: Penguin Books.

DeCoster, J., Iselin, A. M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, *14*(4), 349–366.

Denollet, J., Sys, S. U., & Brutsaert, D. L. (1995). Personality and mortality after myocardial infarction. *Psychosomatic Medicine*, *57*(6), 582–591.

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology*, *60*(6), 817–830.

Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrimination, and racism: Historical trends and contemporary approaches*. Orlando, FL: Academic Press.

Einstein, A. (1920/2001). *Relativity: The special and the general theory* (Reprint of 1920 translation by Robert W. Lawson, ed., p. 48). London, UK: Routledge. ISBN 0-415-25384-5

Einstein, A. (1934). On the method of theoretical physics. *Philosophy of Science*, *1*(2), 163–169.

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *62*, 328–339.

Fallshore, M., & Schooler, J. W. (1993). Post-encoding verbalization impairs transfer on artificial grammar tasks. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 412–416). Erlbaum, Hillsdale, NJ.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, *14*(1), 51–71.

Gehr, B. T., Weiss, C., & Porzsolt, F. (2006). The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Medical Research Methodology*, *6*(1), 25.

Gilbert, G. M. (1951). Stereotype persistence and change among college students. *Journal of Abnormal and Social Psychology*, *46*, 245–254.

Grabe, S., Ward, L. M., & Hyde, J. S. (2008). The role of the media in body image concerns among women: A meta-analysis of experimental and correlational studies. *Psychological Bulletin*, *134*(3), 460.

Greenberg, J., Pyszczynski, T., Solomon, S., Rosenblatt, A., Veeder, M., Kirkland, S., & Lyon, D. (1990). Evidence for terror management II: The effects of mortality salience on reactions to those who threaten or bolster the cultural worldview. *Journal of Personality and Social Psychology*, *58*, 308–318.

Hartley, B. L., & Sutton, R. M. (2013). A stereotype threat account of boys' academic underachievement. *Child Development*, *84*(5), 1716–1733.

Horvath, J. C., Forte, J. D., & Carter, O. (2015). Evidence that transcranial direct current stimulation (tDCS) generates little-to-no reliable neurophysiologic effect beyond MEP amplitude modulation in healthy human subjects: A systematic review. *Neuropsychologia*, *66*, 213–236.

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. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA: The Journal of the American Medical Association*, *279*(4), 281–286.

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA: The Journal of the American Medical Association*, *294*(2), 218–228.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Ioannidis, J. P. A., Polycarpou, A., Ntais, C., & Pavlidis, N. (2003). Randomised trials comparing chemotherapy regimens for advanced non-small cell lung cancer: Biases and evolution over time. *European Journal of Cancer*, *39*(16), 2278–2287.

Ioannidis, J., Trikalinos, T. A., Ntzani, E. E., & Contopoulos-Ioannidis, D. G. (2003). Genetic associations in large versus small studies: an empirical assessment. *The Lancet*, *361*(9357), 567–571.

Ioannidis, J., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549.

Jacobson, J. W., Mulick, J. A., & Schwartz, A. A. (1995). A history of facilitated communication: Science, pseudoscience, and antiscience science working group on facilitated communication. *American Psychologist*, *50*(9), 750.

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 of London. Series B: Biological Sciences*, *269*(1486), 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*(5), 524–532.

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an antidepressive treatment is falling: A meta-analysis. *Psychological Bulletin*. Advance online publication. http://dx.doi.org/10.1037/bul0000015

Karban, R., & Myers, J. H. (1989). Induced plant responses to herbivory. *Annual Review of Ecology and Systematics*, *20*, 331–348.

Karlins, M., Coffman, T. L., & Walters, G. (1969). On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, *13*(1), 1–16.

Katz, D., & Braly, K. W. (1933). Racial stereotypes of one-hundred college students. *Journal of Abnormal and Social Psychology*, *28*, 282–290.

Kayande, U., & Bhargava, M. (1994). An examination of temporal patterns in meta-analysis. *Marketing Letters*, *5*(2), 141–151.

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*(3), 504–509.

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

Leimu, R., & Koricheva, J. (2004). Cumulative meta-analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1551), 1961–1966.

Lindsay, R. C., & Wells, G. L. (1980). What price justice? *Law and Human Behavior*, *4*(4), 303–313.

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, *70*(3), 556.

Madon, S., Guyll, M., Aboufadel, K., Montiel, E., Smith, A., Palumbo, P., & Jussim, L. (2001). Ethnic and national stereotypes: The Princeton trilogy revisited and revised. *Personality and Social Psychology Bulletin*, *27*(8), 996–1010.

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, *66*(4), 482–489.

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

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, *35*(2), 231–245.

Møller, A. P., & Thornhill, R. (1998). Bilateral symmetry and sexual selection: a meta-analysis. *The American Naturalist*, *151*(2), 174–192.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, USA*, *106*, 10593–10597.

Nykänen, H., & Koricheva, J. (2004). Damage-induced changes in woody plants and their effects on insect herbivore performance: A meta-analysis. *Oikos*, *104*(2), 247–268.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLOS ONE*, *7*, e42510. doi:10.1371

Pashler, H., Rohrer, D., & Harris, C. R. (2013). Can the goal of honesty be primed? *Journal of Experimental Social Psychology*, *49*, 959–964.

Pereira, T. V., Horwitz, R. I., & Ioannidis, J. P. (2012). Empirical evaluation of very large treatment effects of medical interventions evaluation of very large treatment effects. *JAMA*, *308*(16), 1676–1684.

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect–Shmozart effect: A meta-analysis. *Intelligence*, *38*(3), 314–323.

Plant, E. A., Devine, P. G., & Brazy, P. C. (2003). The bogus pipeline and motivations to respond without prejudice: Revisiting the fading and faking of racial prejudice. *Group Processes & Intergroup Relations*, *6*(2), 187–200.

Plante, I., Theoret, M., & Favreau, O. E. (2009). Student gender stereotypes: Contrasting the perceived maleness and femaleness of mathematics and language. *Educational Psychology*, *29*(4), 385–405.

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1445), 787–792.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G., Newman, G. E., Wurzacher, O., Nowak, M. A., & Greene, J. D. (2013). Intuitive cooperation and the social heuristics hypothesis: Evidence from 15 time constraint studies. Available at SSRN: http:/ssrn.com/abstract=2222683.

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature*, *365*(6447), 611.

Rideout, B. E., & Taylor, J. (1997). Enhanced spatial performance following 10 minutes exposure to music: A replication. *Perceptual and Motor Skills*, *85*(1), 112–114.

Rosenthal, R. (2005). Experimenter effects. *Encyclopedia of Social Measurement*, *1*, 871–875.

Ryan, R. S., & Schooler, J. W. (1998). Whom do words hurt? Individual differences in susceptibility to verbal overshadowing. *Applied Cognitive Psychology*, *12*, 105–125.

Sánchez, M. I., Georgiev, B. B., & Green, A. J. (2007). Avian cestodes affect the behaviour of their intermediate host *Artemia parthenogenetica*: An experimental study. *Behavioural Processes*, *74*(3), 293–299.

Schlosshauer, M., Kofler, J., & Zeilinger, A. (2013). A snapshot of foundational attitudes toward quantum mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *44*(3), 222–230.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437.

Schooler, J. W. (2014a). Metascience could rescue the "replication crisis." *Nature*, *515*, 9.

Schooler, J. W. (2014b). Turning the lens of science on itself verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science*, *9*(5), 579–584.

Schooler, J. (2015). Bridging the objective/subjective divide – towards a meta-perspective of science and experience. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*: 34(T). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570405

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71.

Schooler, J. W., Fiore, S. M., & Brandimonte, M. (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*(2), 166–183.

Scott, G., Leritz, L. E., & Mumford, M. D. (2004). The effectiveness of creativity training: A quantitative review. *Creativity Research Journal*, *16*(4), 361–388.

Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology*, *18*(2), 247–255.

Simmons, L. W., Tomkins, J. L., Kotiaho, J. S., & Hunt, J. (1999). Fluctuating paradigm. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1419), 593–595.

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*(11), 1359–1366.

Simon, L., Greenberg, J., Harmon-Jones, E., Solomon, S., Pyszczynski, T., Arndt, J., & Abend, T. (1997). Terror management and cognitive-experiential self-theory: Evidence that terror management occurs in the experiential system. *Journal of Personality and Social Psychology*, *72*, 1132–1146.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-Curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Siontis, K. C., Patsopoulos, N. A., & Ioannidis, J. P. (2010). Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics*, *18*(7), 832–837.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Steele, K. M., Bass, K. E., & Crook, M. D. (1999). The mystery of the Mozart effect: Failure to replicate. *Psychological Science*, *10*(4), 366–369.

Sundet, J. M., Tambs, K., Magnus, P., & Berg, K. (1988). On the question of secular trends in the heritability of intelligence test scores: A study of Norwegian twins. *Intelligence*, *12*(1), 47–59.

Swaddle, J. P., Witter, M. S., & Cuthill, I. C. (1994). The analysis of fluctuating asymmetry. *Animal Behaviour*, *48*(4), 986–989.

Tellis, G. J., & Wernerfelt, B. (1987). Competitive price and quality under asymmetric information. *Marketing Science*, *6*(3), 240–253.

Tu, Y. K., Tugnait, A., & Clerehugh, V. (2008). Is there a temporal trend in the reported treatment efficacy of periodontal regeneration? A meta-analysis of randomized-controlled trials. *Journal of Clinical Periodontology*, *35*(2), 139–146.

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*(5), 553–571.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*(10), e201.