Author's Note

This is a preprint of an article published in *Perspectives on Psychological Science* (https://doi.org/10.1177/1745691619838258). As the published version is the "document of record," one should avoid quoting from the preprint and instead refer to the publication.

The substantive content of the two versions does not differ, but the published version has been edited to conform to the journal's style and policies. Notably, *Perspectives* does not allow for abbreviations of citations. Here in the preprint, I use the abbreviation DLMMCC to refer to the meta-analytic review of cues to deception by Bella DePaulo and her colleagues. Figures in the published version appear somewhat different, but the content and underlying data are identical.

Otherwise, differences between the preprint and publication are relatively minor. Any errors in this or the published version are mine.

Timothy J. Luke, June 11, 2019

Lessons from Pinocchio: Cues to deception may be highly exaggerated

Timothy J. Luke*

University of Gothenburg

Abstract

Deception researchers widely acknowledge that cues to deception - observable behaviors that may differ between truthful and deceptive messages - tend to be weak. Nevertheless, several deception cues have been reported with unusually large effect sizes, and some researchers have advocated the use of such cues as tools for detecting deceit and assessing credibility in practical contexts. Examining data from empirical deception cue research and using a series of Monte Carlo simulations, I demonstrate that many estimated effect sizes of deception cues may be greatly inflated by publication bias, small numbers of estimates, and low power. Indeed, simulations indicate the informational value of the present deception literature is quite low, such that it is not possible to determine whether any given effect is real or a false positive. I warn against the hazards of relying on potentially illusory cues to deception and offer some recommendations for improving the state of the science of deception. A preprint of this document is available at https://osf.io/xt8fq/.

> *"That Marionette," continued the Talking Cricket, "is a rascal of the worst kind."* (Collodi, 1883)

In an effort to temper our hopes of catching lies with unerring accuracy, deception researchers often say there is no "Pinocchio's nose" (e.g., Frank, Menasco, & O'Sullivan, 2008; Hartwig & Bond, 2011; Vrij, 2006, 2004). That is, there is no behavior that perfectly discriminates between truthful and deceptive messages. For all the talk about Pinocchio's nose, there are other more urgent lessons from the story of Pinocchio pertinent to deception research.

In his adventures, Pinocchio causes mischief and mayhem, and among his many foibles is the ease with which he is tempted to do what is easy and immediately satisfying but ultimately dangerous, rather than what is difficult and responsible.

*timothy.luke@psy.gu.se. For their helpful comments, suggestions, and inspiration, I am grateful to Fabi Alceste, Karl Ask, Charlie Bond, Will Crozier, Bella DePaulo, Emelie Ernberg, Pär Anders Granhag, Maria Hartwig, Lorraine Hope, Emily Joseph, Louise Jupe, Erik Mac Giolla, Patty Sanchez, Rebecca Willén, and three anonymous reviewers, as well as the many supportive others who provided encouragement and tolerated my ravings and rantings.

Tricksters often get the better of him because they tell him flattering and favorable untruths. Throughout the stories, he is stabbed, hanged from a tree, sold to a circus, and swallowed by a giant fish. All this happens despite the fact that he is warned in advance of nearly every misfortune that befalls him. For the reader who sympathizes with the protagonist, reading these stories is an exercise in frustration: Every time it seems Pinocchio has resolved to do the right thing, he is led astray. It is only after frequent and serious failings that he follows the guidance of the Fairy with Turquoise Hair, corrects his mischievous ways, and becomes "a real boy."

I am concerned deception researchers (and other psychological scientists), like Pinocchio, have ignored good advice. Likely because of a combination of perverse incentives for poor practices and lack of awareness of the consequences of such practices, much research has been conducted in a way that is highly prone to error (Agnoli et al., 2017; Bakker et al., 2012; Ioannadis, 2005; John et al., 2012; Simmons, Nelson, & Simonsohn, 2011). Rather than leading us to being swallowed by a fish, failing to heed available recommendations may have created a grossly distorted view of human deception.

It is widely accepted among researchers that cues to deception are weak (DePaulo et al., 2003; Hartwig & Bond, 2011). Nevertheless, researchers continue to hunt for deception cues and offer practical recommendations for what behaviors to look for in order to more accurately detect lies. For example, Evans and her colleagues (2013), Johnston and his colleagues (2014), and Akehurst and her colleagues (2017) have developed checklists of cues to deception, designed to help practitioners make important decisions about who to believe and who to distrust. In many cases, the inclusion of particular cues in such tools is explicitly justified by the effect sizes for those cues reported in the literature.

The meta-analytic review of cues to deception by DePaulo and her colleagues (2003) found that cues to deception are generally quite weak but also found several cues that significantly distinguished between truthful and deceptive messages, some of which with moderate or large effect sizes. Indeed, DePaulo et al. (2003) is often directly cited both to make the point that cues to deception are weak (e.g., Vrij & Granhag, 2012) and to justify recommendations of examining certain cues to detect deception in practical contexts. For example, Evans and her colleagues (2013) suggest observers should examine, among many other cues, the plausibility of the message (a cue with a meta-analytic estimate of $d = 0.23$ in DePaulo et al., 2003), the amount of detail in the message ($d = 0.30$), and how nervous and tense the speaker is ($d = 0.27$). At face value, it seems that although most behaviors do not distinguish between truthful and deceptive messages, some do – and might be effective for catching lies with reasonable levels of accuracy.

However, this consensus may be incorrect. It is possible that suboptimal research practices have produced a literature rife with false positives. Here, using a series of Monte Carlo simulations, I demonstrate just that: observed effect sizes of deception cues may be greatly inflated by low numbers of estimates, selective reporting, and low power, and in fact, the extant literature is consistent with there being no real cues to deception at all. The informational value of the present literature is so low as to make it virtually impossible to distinguish real effects from false positives.

Lessons from Pinocchio

When a large number of variables are measured to find differences between two groups, small samples and small numbers of estimates can produce the appearance of numerous strong effects even when all effects studied in the literature are in fact zero or negligibly small. This problem arises not only because researchers can selectively report "what works" and infrequently replicate results but also because significant effects are massively inflated when power is low. This is a problem in psychological science at large (Simmons, Nelson, & Simonsohn, 2011; Vul et al., 2009; Yarkoni, 2009). However, features of the deception literature make this problem especially likely. The paradigm in which deception cue data are collected offers unusually high flexibility in coding and analysis, such that a large number of illusory effects can potentially accumulate. The proliferation of uncorrected false positives permits researchers to easily find evidence in the literature for at least some cues to deception that appear to distinguish between truth and deception but in fact may not. Unaware of the error, we can then base further research and practical recommendations on such illusory cues.

How could this problem go on unnoticed and uncorrected? Selective reporting practices and the habit of running low-power studies can be self-reinforcing because they produce what appear to be large effects but are often false positives (Nelson, Simmons, & Simonsohn, 2018). In the face of repeated apparent success, researchers would not necessarily see a need to change their methods (e.g., increase sample sizes) or impose restrictions on their data collection habits or analytic strategies.

As this is essentially a problem of an overabundance of freedom, I call this the *Land of Toys problem*. In the classic story (Collodi, 1883), Pinocchio's journey to becoming a "real boy" takes a long detour when he is tempted by his friend Lamp-Wick to travel to the Land of Toys – a place where children may do as they please and never have to attend school. Although Pinocchio has been repeatedly implored by the Turquoise Fairy and the Talking Cricket not to succumb to laziness, the allure of the Land of Toys is too much for him, and he accompanies Lamp-Wick there. However, Lamp-Wick fails to mention (because he does not know) the high cost of staying too long in the Land of Toys: a magical fever that transforms its victims into donkeys. Pinocchio ends up worse off than he began.

The Land of Toys is a useful analogy because, as we will see, it represents both the causes and consequences of this scientific problem: Researchers deviate from available statistical and methodological recommendations, partly because researchers do not realize how problematic some of their decisions are and partly because there are incentives to engage in questionable research practices. The unintended consequence of this conduct is an outcome antithetical to our objectives – a staggering number of undetected potential false positives. In this analogy, we, deception researchers collectively, are Pinocchio; we have acted in ways that undermine our own goals.

The Land of Toys problem represents a failure (or substantially delayed application) of science's purported feature of self-correction (see Nosek, Spies, & Motyl, 2012). Science is believed to correct its own mistakes over time, as further evidence accumulates to disconfirm previous erroneous conclusions. However, false

positives can live on in a literature, particularly when the evidence typically produced in that literature has a high rate of potential error and when little or no further evidence accrues to correct errors that have occurred (Stroebe, Postmes, & Spears, 2012). Widespread problematic research practices underpin the high error-rate, and when researchers are slow to respond to methodological innovation, the potential for self-correction is especially low.

To illustrate how the Land of Toys problem may be real and present in the deception literature, I draw substantially from the aforementioned widely-cited meta-analytic review of the deception cue literature by DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, and Cooper's (2003; hereafter DLMMCC). Though it is more than a decade old now, DLMMCC remains the most comprehensive review of the deception literature. Here, I use DLMMCC's data as a representation of the deception literature. One can reasonably question the appropriateness of using data from a meta-analyses from so long ago as a representation of the contemporary literature. There are several indications that DLMMCC remains current in numerous ways. First, DLMMCC continues to be cited as an authoritative resource on cues to deception, as can be seen in more recent overviews of the literature (e.g., Hartwig & Bond, 2011, 2014; Bond, Hartwig, & Levine, 2014). Second, as we will see below, research conducted in the years following DLMMCC entails largely the same methodologies (including their shortcomings).

Although I am critical of the DLMMCC's substantive theoretical conclusions and the manner in which deception researchers have used DLMMCC, I do not doubt the rigor with which DLMMCC's data were extracted from the literature. The data I use come principally from two sources: (1) the published version of DLMMCC's review, from which I extracted much quantitative information, and (2) the original data DLMMCC extracted from the literature[1].

Trouble in the Land of Toys

> *The obvious precaution is computation.* - Tversky and Kahneman (1971, p.110)

> *"DOWN WITH ARITHMETIC."* – Graffiti in the Land of Toys (Collodi, 1883)

The deception cue paradigm. Vul and his colleagues (2009) (in)famously noted the reporting of numerous implausibly strong correlations (e.g., $r = |.80|$ and above) in social neuroscience research – colloquially, "voodoo correlations." These correlations appear to be the result of nonindependent analysis and low statistical power (Yarkoni, 2009). One of the features of social neuroscience that supported the reporting of voodoo correlations is that fMRI datasets comprise a

[1]Some years ago, I requested a copy of the data from Bella DePaulo, first author of DLMMCC, as there is a note in the published version indicating the data were available from her. She indicated that she was willing to share the data but no longer had the file. Ultimately, I obtained the original data from Charlie Bond, who is not an author of DLMMCC but who had received a copy of it from Bella DePaulo in order to conduct prior reanalyses (e.g., Bond, Hartwig, & Levine, 2014).

large number of voxels that can be selected, grouped, and analyzed in a variety of ways. This flexibility can lead researchers to perform selective searches for significant correlations. A similar situation exists in deception cue research, although it likely offers somewhat less analytic flexibility than fMRI data.

Deception cue studies produce or obtain records of truthful and deceptive messages. These records can be written, audio, and/or visual. The records are then coded by human raters or automated coding software for variables (i.e., cues) that may be related to deception. Researchers can then perform significance tests and calculate effect sizes measuring the extent to which liars and truth-tellers differ. Observable differences between deceptive and truthful messages have potential theoretical and practical importance, so cues that discriminate between lies and truths are desiderata of deception research. Indeed, because the deception literature is concerned with all potential discernible differences in truthful and deceptive behavior, coding virtually any behavior can be justified, and results in both positive and negative directions are desirable. Research designs often involve manipulations not only of the veracity of messages but also of potential moderators of cues to deception (e.g., preparation, motivation to succeed). As such, there are often numerous groups that can be compared.

Additionally, because researchers are free to choose how to code the messages they have, it is possible to code any number of cues. It is common for researchers to code numerous cues to deception. In DLMMCC, studies on average reported $M = 6.65$ cues ($SD = 5.72$, median = 4, range 1 to 27). If untethered by preregistration or standardized coding, researchers are limited in their ability and flexibility to code data primarily by resource constraints and incentives to report results quickly (rather than spend an indefinite amount of time coding new variables).

In some ways, flexibility is desirable. It provides researchers with the freedom to explore their data thoroughly for interesting patterns and to make unexpected discoveries. But the two demons of virtually limitless flexibility and desire to publish interesting significant results can tempt researchers toward questionable practices (Nosek, Spies, & Motyl, 2012). They might code numerous cues and report only some of them. They might analyze their data in several ways (e.g., subsetting data, collapsing or dropping conditions, excluding participants), until they find favorable results (see Simmons, Nelson, & Simonsohn, 2011).

This kind of problematic flexibility – and the freedom to make numerous undisclosed decisions about data collection, coding, and analysis – is, of course, not unique to the deception literature, but the typical paradigm of deception cue research may provide especially fertile ground for questionable practices. Specifically, the number of measurements deception researchers regularly take and the extensive flexibility in coding decisions exacerbate the risks incurred by problematic research practices that are highly prevalent in psychological science, namely conducting studies with low power and selectively reporting results.

Problematic research practices in the deception literature: Low power, selective reporting, and too many positive results.

> *One might think that after 1969, when I published my power handbook that made power analysis as easy as falling off a log, the concepts and methods of power analysis would be taken to the hearts of null*

> *hypothesis testers. So one might think. (Stay tuned.)* - Jacob Cohen (1990, p.1308)

> *"Don't you know that if you go on like that, you will grow into a perfect donkey and that you'll be the laughingstock of everyone?"* - The Talking Cricket, shortly before Pinocchio kills him with a hammer (Collodi, 1883)

The slowness, if not resistance, of psychological scientists to adopt better statistical and methodological practices has been well-documented (see, e.g., Cohen, 1994, 1990; Cumming et al., 2007; Fidler et al., 2004; Gigerenzer, 2004; Sharpe, 2013). Despite the long existence of an extensive methodological literature identifying poor scientific practices and proposing solutions (e.g., Cohen, 1969, 1962; Meehl, 1978; Sterling, 1959), researchers frequently misunderstand and inadequately address statistical concepts fundamental to their methodologies, such as power (Bakker et al., 2016; Tversky & Kahneman, 1971) and p-values (Gigerenzer, 2004; Greenland et al., 2016). Separately but relatedly, meta-scientists have also documented the alarmingly wide prevalence of questionable research practices in psychology (and other sciences), such as selective reporting, data peeking, unplanned statistical analyses, and hypothesizing after the results are known (see, e.g., Agnoli et al., 2017; Bakker et al., 2012; Fraser et al., 2018; John, Lowenstein, & Prelec, 2012; Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011). There is ample evidence that methodological flaws in psychology and other disciplines have persisted in spite of clear evidence of their occurrence and the existence of productive alternatives. For example, methodological developments and studies thereof throughout past decades have been accompanied by few changes in statistical practice (Cumming et al., 2007) and only nominal increases in sample sizes (Rossi, 1990; Sedlmeier & Gigerenzer, 1989). Although the advice is old, it is only relatively recently that a robust movement toward better practices appears to have taken hold (Cumming, 2014; Nelson, Simmons, & Simonsohn, 2018; Vazire, 2018a).

Does the deception literature share problematic practices with the rest of psychology? Evidence suggests it does. Below, I examine data to assess statistical power, selective reporting, and the rate of reported significant results in the deception literature.

Statistical power. The statistical power of a test refers to the probability of correctly rejecting the null hypothesis, assuming that a true effect exists in the population. Power is a function of the alpha level (i.e., the threshold for significance), the true effect size, and the size of the sample. Because significance levels are generally decided by the conventions of the broader field and because the true effect size is unknown, sample size is the source of power primarily under researchers' control. Achieving adequate power to detect plausible effects is critically important, as it directly determines the informational value of a study's results. Low power can render nonsignificant results uninterpretable (Cohen, 1988; Morey & Lakens, 2016; Tverskey & Kahneman, 1971) and can render significant results untrustworthy (as we will see in more detail later).

To assess power in the deception literature, I drew from two meta-analytic reviews: DLMMCC, which comprehensively reviewed the deception cue literature

from 1920 to 2001[2], and Amado, Arce, Fariña, and Vilariño (2016), which synthesized research on Criterion-Based Content Analysis (CBCA; for an introduction see, e.g., Vrij, 2015), a deception detection approach that measures numerous cues to deception. For a separate project, I recently collected and reanalyzed the literature from Amado et al. (2016; see RabbitSnore, 2018, for a description of how the data were assembled). Their review encompasses literature from 1993 to 2015, and here I examined data for 36 of the studies included in the review that reported individual cues (rather than composites of multiple cues).

In DLMMCC, the average total sample size was $N = 41.18$ ($SD = 31.83$, median = 34); the largest was $N = 192$ and the smallest was $N = 5$. Figure 1 depicts statistical power in the literature reviewed by DLMMCC. The curves on the plot illustrate the largest sample, smallest sample, mean sample, and median sample, with power on the vertical axis and effect size on the horizontal axis. One can see, for example, that the largest sample in DLMMCC had approximately .80 power to detect an effect of 0.40 – and all other studies in the literature had less power.

The median effect size (of effect sizes with at least three estimates) found in DLMMCC was $d = |0.10|$, for which the included studies were highly underpowered. The average included study had .06 power to detect an effect of this size (assuming a two-tailed test). The largest included study had less than .11 power. That is, the largest deception study would fail to detect the median deception cue, as estimated by DLMMCC, roughly 9 times out of 10. The largest effect (from at least three estimates) in DLMMCC was $d = 0.66$. The average included study had .55 power to detect an effect of this size. Power in the deception literature has been extremely low[3].

Figure 2 depicts statistical power in the literature reviewed by Amado et al. (2016). Sample sizes in the CBCA literature have tended to be larger than earlier deception studies, with an average sample size of $N = 76.4$ ($SD = 66.4$, median = 60.0). This average may be disproportionately affected by three outlying samples with $N > 250$ (two of which are unpublished). Removing these outliers, the average sample size is $N = 58.2$ ($SD = 24.2$, median = 59.0) – still larger than the literature reviewed in DLMMCC. The power of this literature is, however, still relatively unimpressive (assuming $N = 76$, .93 power for $d = 0.80$, .73 for $d = 0.60$, .57 for $d = 0.50$, .14 for $d = 0.20$), given that cues to deception are likely to have quite modest effect sizes. The average study reviewed in Amado et al. (2016) had .80 power to detect $d = 0.65$ – a massive effect compared to typically observed deception cue effects.

Sample sizes in the deception literature appear to have increased somewhat in recent years. This has been the case in other domains of psychological science as well. But in the words of Rossi (1990), "these increases are no cause for joy"

[2]Of the 142 samples, data for 139 were published in 1970 or later. The other three were published in 1920, 1923, and 1943.

[3]To borrow a reference point from Simmons, Nelson, and Simonsohn (2013, 2018), the effect size for the difference in weight between men and women is about $d = 0.60$. The average study in DLMMCC (assuming $N = 41$) had .47 power to detect an effect of this size. Thus, the literature has not only been underpowered to detect effects that we now know are plausible, but it has also been so underpowered that the average study would, more than half the time, fail to find a difference that can be detected by casual observation in everyday life.

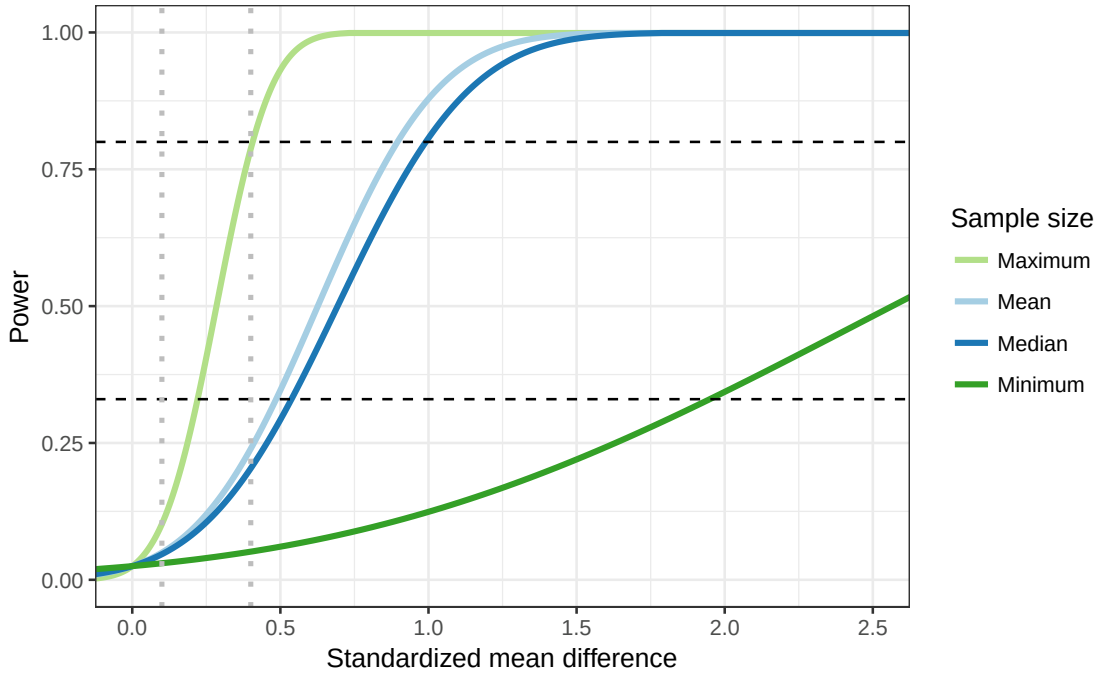

Figure 1. *Statistical power in DLMMCC, years 1920 to 2001*
Each curve represents a different sample size from DLMMCC (i.e., the largest sample, mean sample, median sample, and smallest sample). Vertical dotted lines are drawn at $d = 0.10$ (the median cue effect in DLMMCC) and 0.40 (approximately the average effect in social psychology; Richard, Bond, Stokes-Zoota, 2003). Horizontal dashed lines are drawn at .33 power and .80 power. Examining the curves, one can see how much power that sample had to detect effects of a given size. For example, the median sample had less than .25 power to detect an effect of $d = 0.40$. Resources to reproduce this figure can be found at https://osf.io/mfq6u/.

(p.650). Power to detect plausible effects in the deception cue literature has been low and continues to be low.

Selective reporting. The term "selective reporting" can refer to a variety of questionable practices in which researchers report some but not all of their results, measurements, conditions, or analytic decisions. Sometimes entire studies are not reported (Rosenthal, 1979), but selective reporting is a problem in published work as well, as flexibility in the disclosure of results and methods can greatly increase the risk of false positives (Simmons, Nelson, & Simonsohn, 2011). In accumulation, selective reporting cripples a field's ability to correct its own errors (Bakker et al., 2012; Ioannidis, 2005; Nosek, Spies, & Motyl, 2012). However, surveys of psychological scientists and researchers in other fields find that selective reporting is highly prevalent (Agnoli et al., 2017; Fraser et al., 2018; John et al., 2012; Martinson et al., 2005).

By definition, we do not know how many estimates have gone unreported in the deception literature, but there are signs in the literature that there is extensive selective reporting: When nonsignficant cues are reported, they are often only reported in such a way that it is not possible to calculate a precise effect size. DLMMCC report that of the 1,338 effect sizes in the literature, 787 (58.8%) could

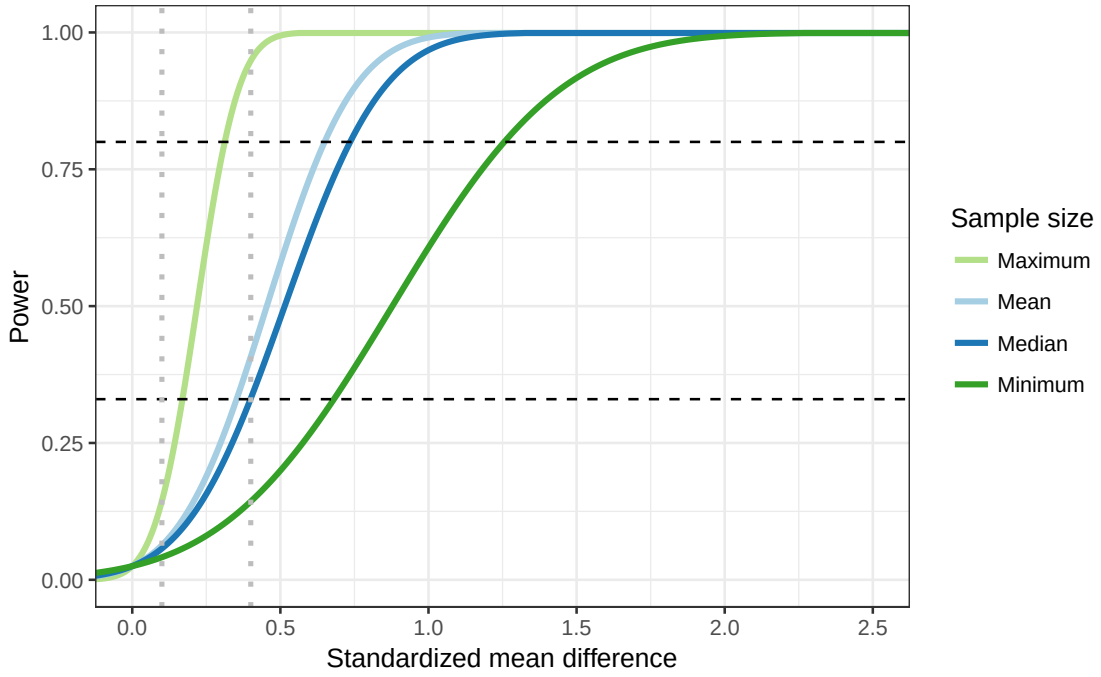

Figure 2. *Statistical power in in Amado et al. (2016), years 1993 to 2015*
Each curve represents a different sample size from Amado et al. (2016) (i.e., the largest sample, mean sample, median sample, and smallest sample). Vertical dotted lines are drawn at $d = 0.10$ and 0.40. Horizontal dashed lines are drawn at .33 power and .80 power. Resources to reproduce this figure can be found at https://osf.io/mfq6u/.

be extracted precisely from the documents. By simple arithmetic, 551 (41.2%) of the effects were reported minimally (e.g., reported as nonsignificant with little supporting information). That is, we are missing about 2 effects out of every 5. These are, of course, the effects that actually made it into reports; presumably there are many more estimates that have been lost to the file drawer (see, e.g., Easterbrook et al., 1991; Fanelli, 2012; and see Appendix A).

Interestingly, selective reporting is far more prevalent in published studies (45.1%, 544 out of 1206 effects) compared to unpublished studies included in DLMMCC (5.3%, 7 out of 132 effects). One possible explanation for this striking pattern is that selective reporting is a strategic decision made to facilitate publication (e.g., downplaying nonsignificant results to bolster credibility; removing "uninteresting" results at the suggestion of reviewers). In contrast, unpublished work such as theses and dissertations often have much more liberty to be transparent (see Mazzola & Deuling, 2013), as there are fewer space restrictions and their "success" may depend less on statistical significance than published papers (Bakker et al., 2012; Easterbrook, 1991; Greenwald, 1975). The majority of unpublished material included in DLMMCC were theses and dissertations.

For more evidence of selective reporting, we can also look to the data from Amado et al. (2016). When I (re)collected those data, whenever it was not possible to calculate an exact effect size, I recorded the reason (if one was stated or discernible) the data were not reported. Data were frequently unreported because

there was no significant difference between truthful and deceptive statements, sometimes related to restriction of range in the cue. In such situations, best practices dictate reporting sufficient statistics to facilitate a meta-analysis, so failure to report the necessary data represents a kind of problematic selective reporting. Thus, one way to assess selective reporting in the CBCA literature is to examine the proportion of such unreported effects, out of the total number of measured effects (minus those not reported for other reasons). Calculated this way, out of $k = 432$ effects in the CBCA literature, 80 of them (18.5%) were excluded from reports. Here, we are missing approximately 1 effect out of every 5.

A word of caution: We do not know how accurate of an estimate of selectively reporting dependent variables these metrics provide. Additionally, these methods provide no information about other kinds of selective reporting (e.g., dropping conditions). However, it is clear that selective reporting is prevalent in the deception literature.

An overabundance of significant results. Yet another potential indicator of selective reporting (or other questionable practices) is the amount of significant results in the published literature. When a literature produces more significant results than there is apparent power to actually obtain, it is a sign of problematic practices (see, e.g., Nelson, Simmons, & Simonsohn, 2018). Using the extracted effect sizes and sample sizes in the original data of DLMMCC, I calculated p-values for each effect (assuming a Student's t-test). These calculated p-values allow us to conduct several descriptive analyses and power calculations to assess the health of the deception literature.

Of the 787 effect sizes that could be precisely extracted from the literature, 129 of them were significant ($p < .05$, two-tailed). This is approximately 16.3%. On its face, this seems like an unimpressive percentage – but it is actually oddly high, given the statistical power of the literature. Within the subset of the literature from which exact effect sizes could be extracted ($k = 98$ studies), the average total sample size was $N = 46.6$ ($SD = 37.8$, median = 35.5). With this information, we can conduct a sensitivity analysis to find the effect size for which typical deception studies are powered at the level at which they have actually found effects (Cohen, 1988). That is, we can assume that the rate of significant results corresponds to the studies' power – the rate at which the null hypothesis is correctly rejected. Assuming a sample size of $N = 46$, the typical deception study would have .163 power to detect a true effect of $d = 0.29$. That is, if cues typically had effects of about 0.29, we would expect to obtain this rate of significant results. However, given that cues' meta-analytically estimated effects are considerably smaller (see Appendix B), the rate of significant effects in the deception literature is remarkably high [4].

Alternatively, one could examine the data at the level of the study. The 98 studies from which exact effects could be extracted reported an average of $M =$

[4] A similarly unusually high rate of significant results is present in the more recent CBCA literature reviewed by Amado and her colleagues (2016), in which 36.2% of effects are significant. Using the mean sample size of $N = 76$ and .362 power, we would expect a typical effect size of $d = 0.37$. The average meta-analytic estimate for cues in Amado et al. (2016) was $d = .25$. It appears the problem of overabundant significant results has not dissipated in the years since DLMMCC was published.

8.03 effects (both significant and nonsignificant; $SD = 8.96$, median = 4), and within each study, on average 21.6% ($SD = 33.3\%$) of effects were significant. We can use this rate of significant effects for a power calculation, as above. Assuming a sample size of $N = 46$, the typical deception study would have .216 power to detect a true effect of $d = 0.35$. Again, given the apparent typical size of deception cues' effects, this rate of significant results is unusually high.

According to my reconstructed p-values, of the 98 studies from which exact effects could be extracted, 43 of them (43.8%) reported at least one significant effect for a cue ($M = 1.32$ significant effects, $SD = 2.46$). At face value, this rate might inspire some confidence that the deception literature is marked by unusual transparency in the reporting of negative results. However, a closer examination of studies that apparently reported no significant cues tells a different story.

There were 55 (out of 98; 56.2%) studies that did not appear to contain significant cue effects. I obtained the 48 that were published and compared their results to my reconstructed p-values. Of these 48, 46 (95.8%) of them reported at least one significant effect of some kind (not always related to deception cues)[5], and 43 (89.6%) of them presented significant results specifically for at least one deception cue. This means at least 86 out of 98 (87.8%)[6] studies from which precise estimates could be extracted presented at least one deception cue as significant.

Why are the rates of reconstructed significant results and reported significant results so widely discrepant? There are a variety of reasons. Sometimes, authors reported finding significant results only in subsets of their data (e.g., Anolli & Ciceri, 1997; Burns & Kintz, 1976; Ekman, Friesen & Simons, 1985). Some authors reported significance for multivariate tests that analyzed cues together (e.g., Greene et al., 1985; Heilveil & Muehleman, 1981) or analyses that collapsed across other conditions (e.g., DePaulo, Rosenthal, Green, & Rosenkrantz, 1982). These are practices known to inflate the false positive rate (Simmons, Nelson, & Simonsoh, 2011). Some reported statistics that simply appear incorrect (e.g., Heilveil, 1976, reported t [11] = 2.03 as significant though it does not actually reach the critical value). I point this out not to accuse the cited authors or others specifically of malfeasance but rather to account for the discrepancy and to note what appear to have been common analytic practices in the literature.

How can the rate of significant results be so high, given the statistical power of the deception literature? The tendency to favor reporting and publication of positive results is a general problem in science (see, e.g., Fanelli, 2012). Across disciplines, more significant effects have been reported than there has apparently been power to find (Button et al., 2013). There are many potential explanations for this. The classic explanation is, of course, that some studies (or some estimates within studies) are stuffed in the proverbial file drawer and hidden from view (Rosenthal, 1979). However, other prevalent questionable practices, such as data peeking or *ad hoc* dropping of participants, can also lead to inflated rates of positive results (Bakker et al, 2012; Simmons, Nelson, & Simonsohn, 2011). In-

[5]The exceptions were Marston (1920) and Goldstein (1923), the earliest deception cue studies in the literature. Marston and Goldstein did not use p-values, so they could not present significant results.

[6]A data file recording which studies I noted to report significant results is available here: https://osf.io/mfq6u/

deed, Nelson, Simmons, and Simonsohn (2018) argue that such practices are "the only... practical way to *consistently* get underpowered studies to be statistically significant" (p.515, emphasis in the original).

As I have described previously, the deception cue research paradigm provides for substantial flexibility in data collection and coding, providing especially fertile ground for false positives to grow if deception researchers have engaged in questionable practices as in the rest of psychology (Agnoli et al., 2017; John et al., 2012). For instance, if researchers make *post hoc* decisions to code more deception cues after checking their results, this capitalization on chance could lead to an inflated rate of significant results. Such practices can have effects similar to conventional publication bias (Bakker et al, 2012; Phillips, 2004). In many subject areas, researchers are limited in the dependent variables they can analyze by what they measured at the time they collected the data. Because deception researchers are often able to code more variables, the deception research provides unusually ample opportunities for this kind of practice. The high rate of significant results is not direct evidence of such behavior. However, it would be difficult to produce such high rates of significant results without questionable research practices or stupendous luck (Nelson, Simmons, & Simonsohn, 2018).

Summary. The deception literature exhibits many of the same problems of psychological science at large, and there is no evidence that these problems have been effectively solved since DLMMCC was published many years ago. Rather than being uniquely poor in statistics and methods, what makes the deception cue literature particularly problematic is the paradigm in which the work is conducted *in combination* with methodological flaws more broadly present in psychological science. In other words, in the deception literature, we face the usual hazards and more.

The consequences. Selective reporting can lead to considerable exaggeration of effects, particularly when studies tend to be low-powered. If one obtains a significant p-value and the true effect size is smaller than the study is properly powered to detect, the estimate is necessarily inflated. This occurs, quite simply, because the only way small samples can obtain significant p-values is by observing relatively large effects (Gelman & Carlin, 2014; Lane & Dunlap, 1978; Schmidt, 1992; Yarkoni, 2009). Figure 3 illustrates the results of a Monte Carlo simulation demonstrating how significant effect sizes at low power are highly inflated (see Yarkoni, 2009). The smaller the sample, the greater the inflation. Under such conditions, it is easy to mistake an effect that is actually negligible for a quite substantial one. For example, if you look at the lowest curve on the plot (corresponding to a true effect of 0.10), at $N = 100$, significant results will on average overestimate the effect as approximately 0.40 – inflation by 300%.

One could justifiably object that this simulation does not represent the deception literature and is unreasonably pessimistic. Wouldn't these inflated estimates be reined in by inclusion of nonsignificant results in a meta-analysis? Yes, if every nonsignificant effect were reported and included, a meta-analysis would indeed provide an accurate estimate in the long run. It is not likely, however, that all nonsignificant effects have been reported in the literature. Preferential reporting of significant effects may inflate long run estimates, when some but not all nonsignificant effects are reported.

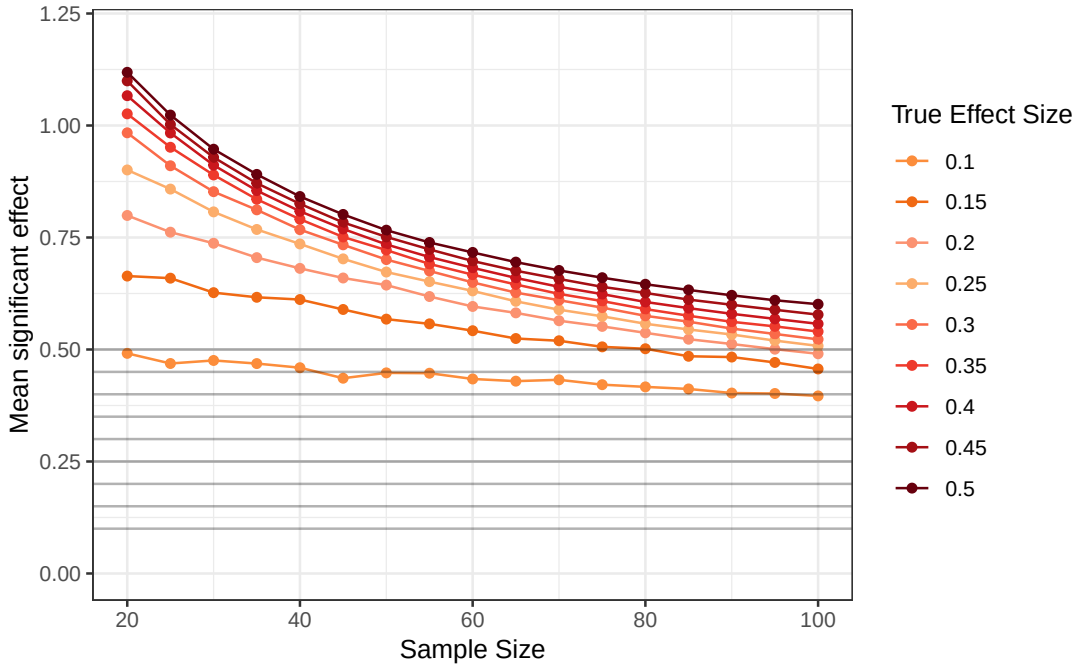

Figure 3. *Inflation of significant effects at low power*
Note: Each point represents the average significant effect ($p < .05$) found in 100,000 simulated studies. Darker horizontal lines are placed at the level of each true effect size. Resources to reproduce this simulation and figure can be found at https://osf.io/gfhqe/.

To illustrate the extent to which publication bias could inflate long run estimates of effects, I conducted a simulation in which I varied the probability of significant and nonsignificant effects being included in the estimate. This simulation was not intended to model the deception literature specifically; rather, it is intended simply as a demonstration of how different levels of bias can produce different long run results. To reflect varying severity of bias against nonsignificant results, I simulated conditions under which 10% to 60% (in intervals of 10%) of nonsignificant effects are included. Perhaps some significant effects are excluded from the literature. To account for this, I simulated conditions under which 80%, 90%, and 100% of significant effects are included[7]. I also varied the sample sizes of each simulated study, from $N = 20$ to 100 in intervals of 5, and the true size of the effect, from $d = 0.10$ to 0.60 at intervals of 0.10. Figure 4 displays the results of this simulation.

[7]I also simulated conditions under which only 70% of significant effects are included, but because this strikes me as an improbably low rate, I excluded it here to conserve space. Data for these simulations are available here: osf.io/gfhqe.

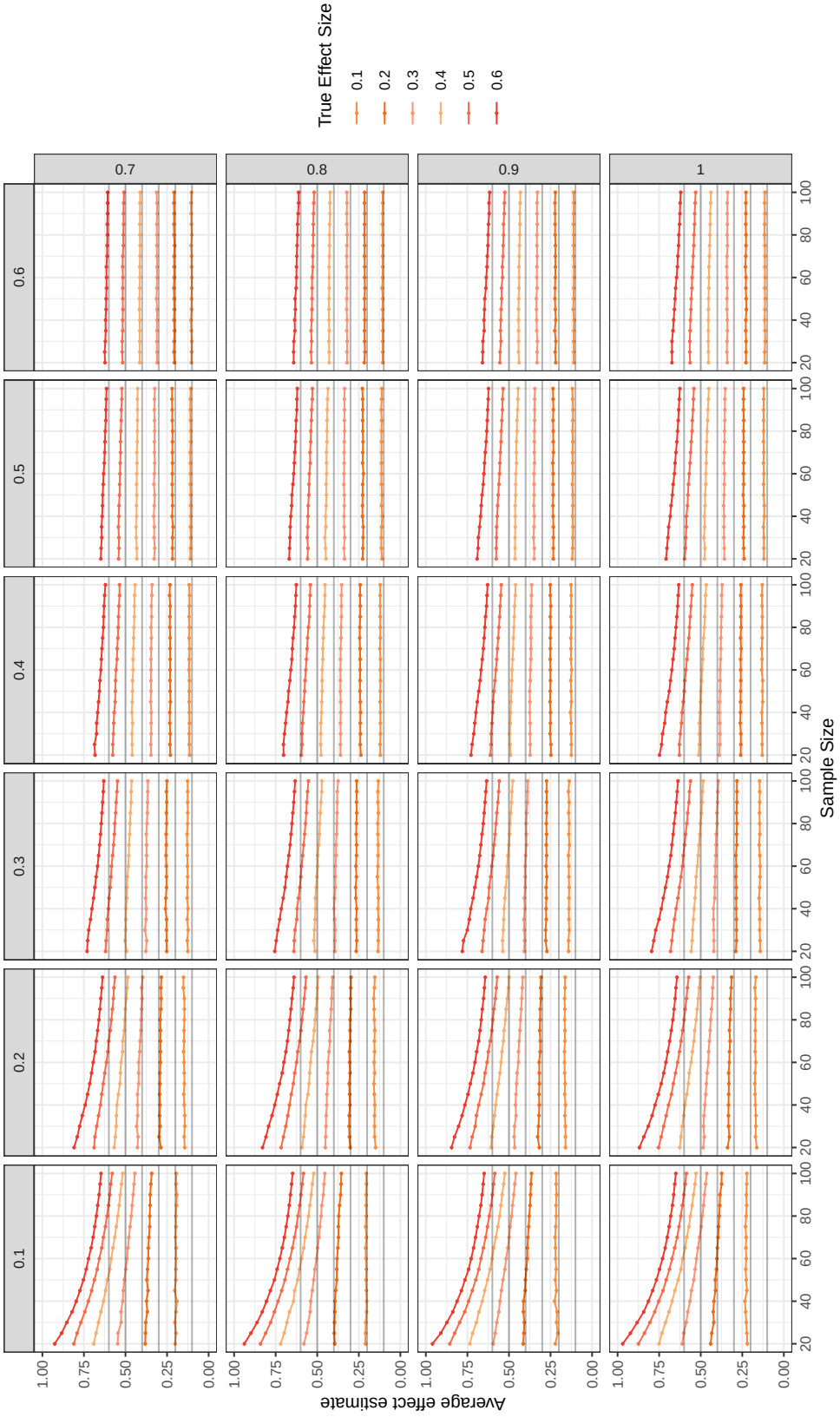

Figure 4. *Long run effect size estimate inflation at low power, with varying severity of publication bias*
Note: Each column of panels represents a different proportion of included nonsignificant results ($p > .05$). Each row of panels represents a different proportion of included significant results ($p < .05$). Each point is estimated with 100,000 simulated studies. Darker horizontal lines are placed at the level of each true effect size. Resources to reproduce this simulation and figure can be found at https://osf.io/gfhqe/.

Without knowing the rate at which estimates are buried in the file drawer, we cannot easily approximate the degree to which effect estimates of deception cues are inflated. However, we can see in Figure 4 that under some circumstances, inflation is severe. As the proportion of included nonsignificant results increases, inflation decreases. Estimates are substantially distorted when only a small proportion of nonsignificant effects are included. However, low power has a greater impact, compared to publication bias, on the precision of estimates (not that it should provide any comfort).

The inflation illustrated in Figure 4 is not a problem unique to the deception literature. Additionally, because it only portrays long run estimates (i.e., derived from large numbers of studies), it does not illustrate the sampling variation that occurs in individual studies, and it does not capture important features of the deception literature. The preceding simulations represent estimates that might obtain when studying a single phenomenon in a large number of studies. As I have previously described, however, deception research is not the continual study of a single phenomenon but rather dozens of cues. Moreover, the number of reported estimates (k) for each cue varies widely (see Figure 5). Sustained and repeated study of individual cues is the exception and not the rule.

Each point in Figure 5 represents a meta-analytic estimate for a cue (except when a cue was only reported once, in which case, it is the raw estimate from that study). Sample sizes and the number of estimates for each cue are plotted on the vertical axis, and the effect size is plotted on the horizontal axis. Thus, we can see how the size of each meta-analytic effect relates to how much it has been studied.

On average, the 158 cues in DLMMCC were reported in $M = 5.94$ studies ($SD = 7.42$, median = 4). The most studied cue was reported $k = 49$ times, and 42 cues were reported only once. It is implausible that this variation in ks is purely the result of publication bias. Rather, it is also likely the case that deception researchers have simply studied some cues more than others.

Additionally, as can be seen in Figure 5, the more a cue has appeared in the literature, the smaller the effect estimate tends to be (see Bond, Levine, and Hartwig, 2014). Indeed, one could be forgiven for mistaking the plots in Figure 5 for funnel plots of estimates homogeneously and fairly symmetrically distributed around zero. But those are not funnel plots. In a meta-analysis of a single homogeneous effect, the shape of a funnel plot is a direct consequence of estimates with higher precision being closer to the true effect size and estimates with lower precision being more widely spread around the true effect size (Duval & Tweedie, 2000). Here, there is no *a priori* reason to expect the distributions in Figure 5 to have such a shape, given that they plot 158 different effects, rather than a single presumably homogeneous effect. If anything, when plotting dozens of potentially diverse effects, one could plausibly expect the estimates to be broadly distributed across the entire plot[8]. Instead, Figure 5 depicts distributions that conform to the shape and location that is expected when a meta-analyzed literature has studied

[8]To illustrate a starkly different pattern from the one in Figure 5, I plotted data from the widely-cited meta-meta-analysis by Richard, Bond, and Stokes-Zoota (2003), which examined 474 meta-analyses in social psychology: https://osf.io/gfhqe/. It is easy to see that these diverse effects, unlike those in the deception cue literature, do not conform to a funnel-like distribution – nor should they.

a truly null or negligible effect.

One can easily see in Figure 5 that large effects only appear for cues reported a small number of times and with smaller total Ns. The number of reports k and total N are both negatively correlated with the absolute value of the cue's effect size, r = -.27 and r = -.26 respectively[9]. Using a simple linear regression approach, the number of studies reporting each cue is negatively related to the absolute value of the cue's effect size, b = -0.007, t (86) = 3.36, $p < .001$. Stated differently, for every additional study reporting a given cue, that cue's effect size is reduced by approximately Δd = -0.007. This may not seem like a dramatic decline, but recall that deception cue studies have on average a sample size of only $N = 41.18$, and cues' effect sizes already tend to be quite modest in size (see Appendix B). Consistent with this trend, total N for each cue is negatively related to the absolute value of the cue's effect size, b = -0.00018, t (86) = 3.53, $p < .001$. That is, for every $N = 100$ participants added to a cue's literature, its effect size tends to shrink by approximately Δd = -0.018.

Researchers have previously offered statistical descriptions of the diminishment of effect sizes over time (Bond, Levine, and Hartwig, 2014), but they have not offered an explanation as to why it occurred. The present analyses, however, suggest this decline effect may be a result of low power and selective reporting: Cues that have rarely been reported will demonstrate wide variation in effect sizes, some of which will appear to be quite large because low-powered studies occasionally produce highly overestimated significant effects, which are more likely to be published than the nonsignificant effects that would attenuate the overestimates (IntHout et al., 2015). In contrast, cues that have been studied more extensively should better approximate the true effect. The shape of the empirical distribution will be influenced by publication bias and heterogeneity. For these reasons, it may be possible to obtain an empirical literature like the one we have even if all the effects were much smaller than their present estimates suggest (or were nonexistent). Under ordinary meta-analytic conditions, we might use publication bias correction techniques (e.g., trim and fill) to obtain better estimates, but for reasons documented in Appendix A, these approaches are limited in their usefulness here. Thus, to examine the possibility that cues to deception may be exaggerated, I conducted another simulation, to which we turn now.

The empirical literature could have been obtained even if every cue's effect size is actually zero. I simulated deception cue literatures that would accumulate under 30 different conditions. Each simulated literature comprises 50,000 cues[10]. Each cue is studied a number of times randomly drawn from a distribution of values that resembles the distribution of ks in DLMMCC. Each study had a sample size randomly drawn from a distribution of values that resembles the distribution of Ns in DLMMCC. The distributions of ks and Ns were identical for each of the simulated literatures.

The reported ks and Ns of the deception literature are known. The distribution of true effect sizes and the severity of the publication bias, however, are unknown.

[9]In an earlier review by Zuckerman, DePaulo, and Rosenthal (1981), there was also a negative correlation between the absolute value of cues' effect sizes and the number of reports, r = -.67.

[10]For details on the simulations, I have archived a more technical and verbose description, which accompanies the simulation code here: https://osf.io/cf5vs/

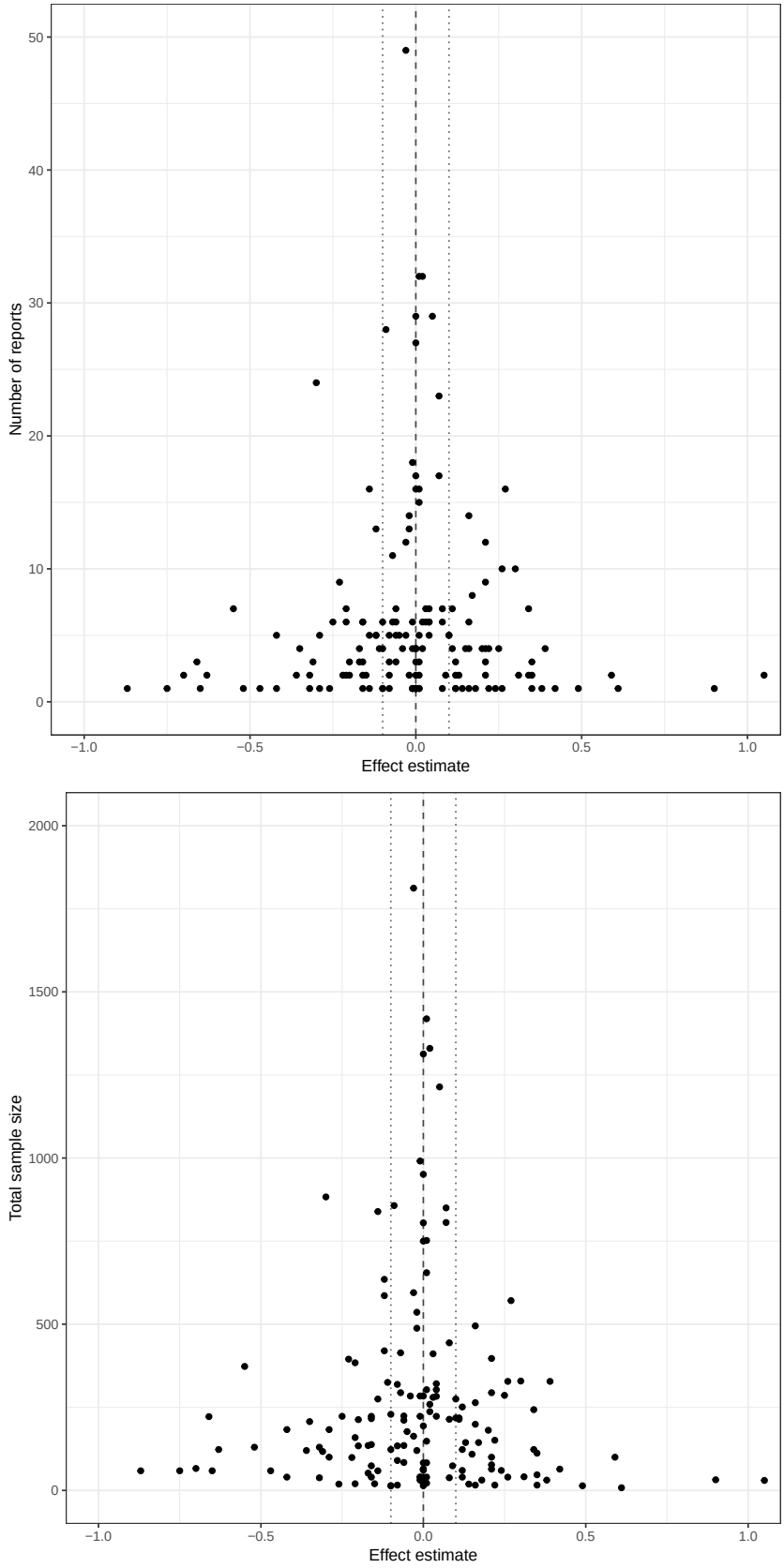

Figure 5. *Estimated effect size by number of studies measuring each cue and total sample size in DLMMCC*

Note: Vertical lines are drawn at d = -0.10, 0, and 0.10. Resources to reproduce this figure can be found at https://osf.io/mfq6u.

To examine different populations of effects from which the empirical literature might be drawn, I created five populations of true effect sizes: (1) all effects equal to $d = 0$; (2) all effects with an absolute value of $d = 0.10$, randomly positive or negative; (3) all effects with an absolute value of $d = 0.25$, randomly positive or negative; (4) effects sampled from a normal distribution with a mean of $d = 0$ and standard deviation of 0.25; and (5) effects sampled from a normal distribution with a mean of $d = 0$ and standard deviation of 0.50. The null population, |.10| population, and |.25| population were included as approximations of a range of plausible effects, given the general pattern of results found by DLMMCC. The population with a mean of $d = 0$ and $SD = 0.25$ reflects a face-value reading of the literature. The mean effect size across the empirical literature is close to 0, and its *SD* is close to 0.25. These parameters can be seen as an assumption that the literature has estimated the effects reported in DLMMCC with a high degree of accuracy. The population with a mean of $d = 0$ and $SD = 0.50$ is included as an illustrative straw-man. It includes more heterogeneity than has actually been observed in the empirical literature, and as such, it represents an unlikely universe in which there are a very substantial number of strong cues to deception. I included this population in the simulations to demonstrate what would happen under such conditions, but I have no illusions about it being plausible.

To model publication bias of different magnitudes, I varied the proportion of nonsignificant effects included in each estimate at six levels: 1.0, .80, .66, .50, .33, and .10. All significant effects were included.

In total, these 30 simulated literatures comprised the results of k = 25,602,957 simulated studies (k = 15,983,427 after exclusions by publication bias) involving N = 1,228,069,423 simulated participants (N = 776,906,246 after publication bias), examining a total of 1,500,000 simulated cues to deception (1,418,689 after publication bias).

Understanding the simulation. What can this simulation tell us? It can provide us a peek at 30 "alternate realities" with known parameters, for comparison to the extant literature. We can examine the outcomes and consider (1) the extent to which each of the parameters is plausible and (2) the extent to which each simulated literature resembles the empirical literature. If a simulated literature closely resembles the literature we actually have, we might infer that the empirical evidence is compatible with the corresponding parameters (assuming they are plausible). Another way to think of such "matching" is that the empirical literature has not provided sufficient evidence to disconfirm the parameters of the matched simulated literature. Because the ks and Ns resemble the empirical literature and we know the true population effect sizes in the simulations, we can assess how accurate the estimates in the empirical literature would be if the specified simulated effects were real. Thus, scrutinizing the way data would have accumulated under specified conditions can tell us extent of the empirical literature's compatibility with various parameters.

Additionally, one can examine the similarity of the simulated literature to each other. If literatures drawn from different populations of true effects (e.g., null vs. |0.10|) are highly similar to each other, it indicates that the precision of the estimates is insufficient to distinguish between those scenarios. If effect populations produce noticeably distinct distributions, we can infer there is sufficient precision

to distinguish between those scenarios. Precision is fixed by the ks and Ns, so we can thus draw conclusions about the precision of the empirical literature by comparing the simulated literatures.

Results of the simulation. Figures 6 and 7 illustrate the results of this simulation. Figure 6 displays effect sizes plotted against the number of studies estimating each cue. Each column of plots belongs to one of the five populations of effect sizes. Each row of plots belongs to one of the six rates of publication bias. Figure 7 displays effect sizes plotted against the total sample size for each cue (viz. the sum of the sample sizes of all studies reporting that cue). Because k and N are naturally highly correlated, these two figures present essentially the same information in different ways. These figures display data in the same way as Figure 5, but because there are so many estimates displayed simultaneously, rather than points, red areas represent the density of the several thousand meta-analytic effect estimates of the simulated cues (darker areas are higher density). For comparison, layered over each simulated literature are the empirical estimates from DLMMCC represented as blue circles. For a closer look, the reader can find higher resolution images of each simulated literature here: https://osf.io/gfhqe/

It is easy to see that the empirical distributions are highly dissimilar to the simulated literatures drawn from the normally distributed effect populations, which are much more widely distributed, even at higher ks and Ns. Even a relatively small amount of variance in true effects would produce a literature radically different from the one we actually have. This is because there is wide variation in the true effects, and there is not sufficient precision in the estimates. Low power leads to wildly inaccurate estimates, so these distributions are exploding with far more variation than we see in the empirical literature. Therefore, if we assume the the true effects have a mean close to 0 and SD close to .25 (as the empirical estimates do), it is extremely unlikely they could have been estimated accurately with these ks and Ns.

Notice that many of the simulations reproduce the decline effect observed in DLMMCC. That is, simulated literatures whose true effects are small demonstrate wide variation at low ks and Ns. This flaring of the distribution of effects is particularly pronounced at moderate and high levels of publication bias. Again, this is due to the tendency for publication bias to select significant effects, which are invariably overestimated at low power. It requires many individual studies to correct for this tendency when an entire literature is underpowered, and even then, there remains substantial error. Cues studied with high ks under conditions of higher publication bias are relatively uncommon (or outright absent), presumably because the bias excludes many of the nonsignificant findings that would have increased the ks and Ns and thus corrected many, though not all, of the overestimates. When publication bias is high, power is low, and true effects are small, the long run is cut short.

The empirical literature is strikingly similar to the simulated literatures drawn from the zero effect and $|0.10|$ populations, with moderate and low publication bias (i.e., the lower left part of Figures 6 and 7; and see Appendix C). Despite the fact that $d = |0.10|$ is the median effect size observed in DLMMCC, the simulated literatures drawn from a population in which all effects are $d = |0.10|$ seem to be more widely distributed at higher ks than the empirical distribution. Indeed,

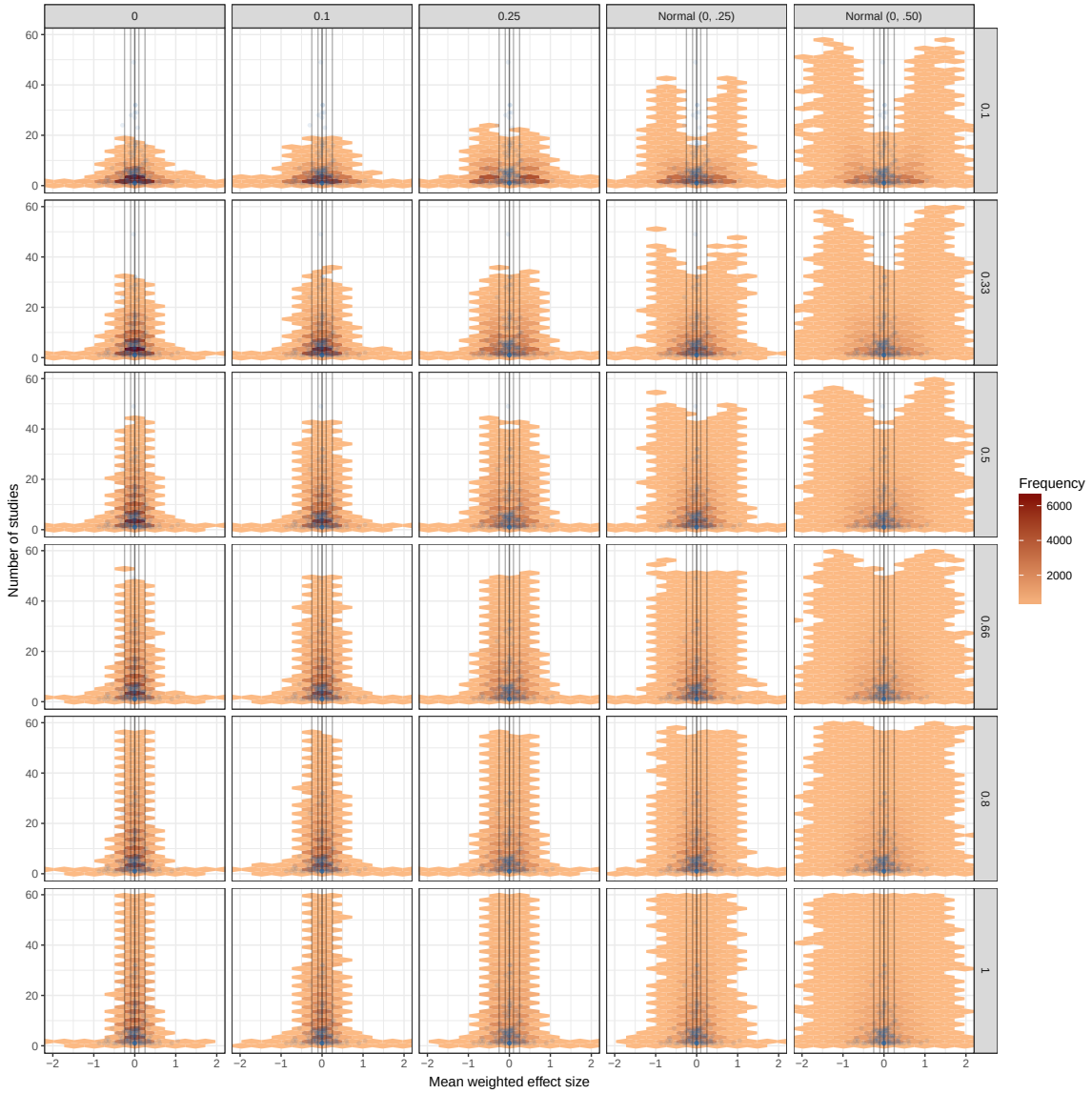

Figure 6. *Simulated deception cue literatures (number of studies)*
Note: Each column of panels represents a population of effect sizes. Each row represents a proportion of nonsignificant effects ($p > .05$) included in the estimates. Red areas represent the density of cue estimates at a given location. Blue circles represent the effect estimates from DLMMCC. Vertical lines are drawn at d = -0.25, -0.10, 0, 0.10, and 0.25. Resources to reproduce this simulation and figure can be found at https://osf.io/gfhqe/.

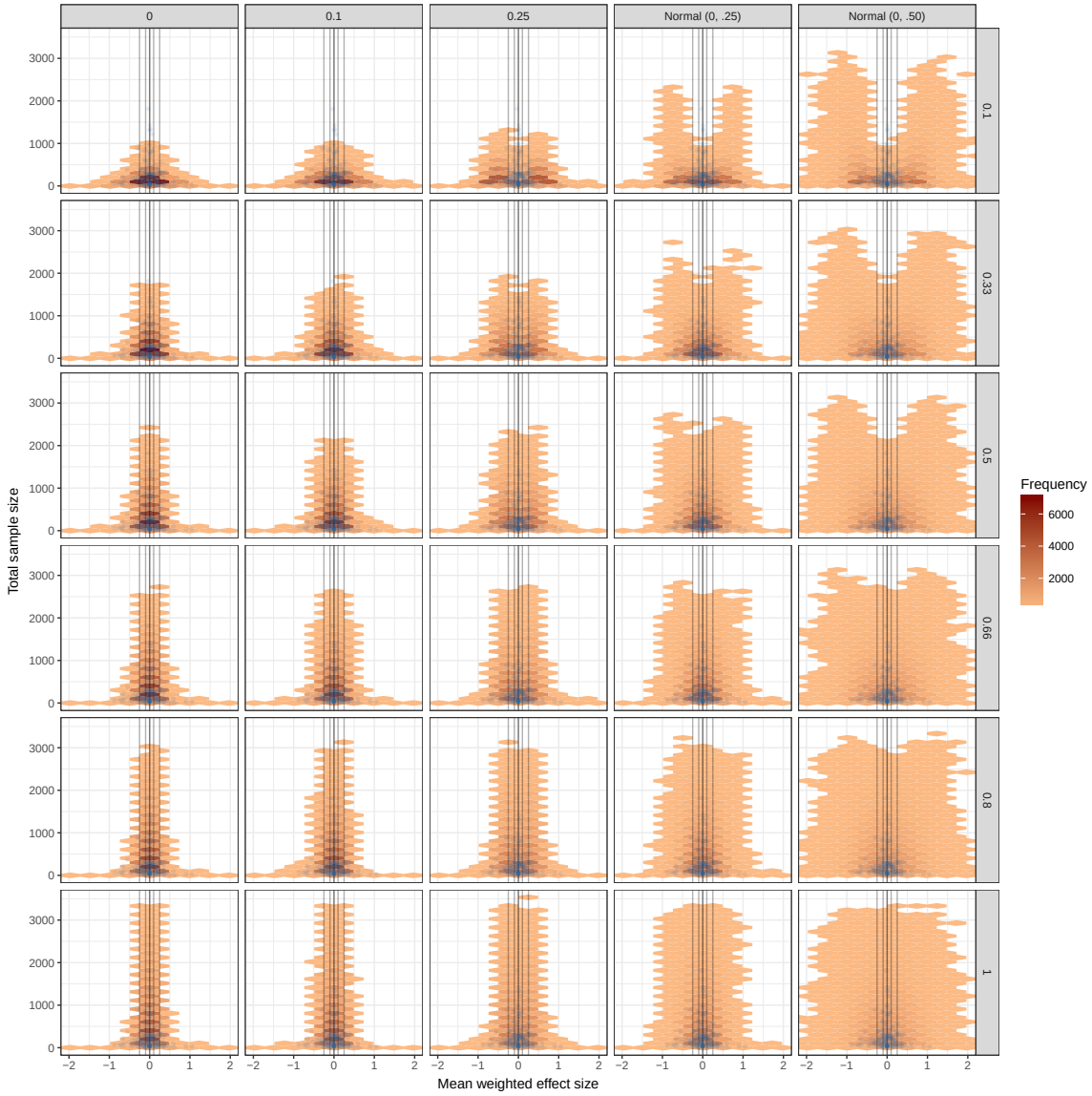

Figure 7. *Simulated deception cue literatures (total sample size)*
Note: Each column of panels represents a population of effect sizes. Each row represents a proportion of nonsignificant effects ($p > .05$) included in the estimates. Red areas represent the density of cue estimates at a given location. Blue circles represent the effect estimates from DLMMCC. Vertical lines are drawn at d = -0.25, -0.10, 0, 0.10, and 0.25. Resources to reproduce this simulation and figure can be found at https://osf.io/gfhqe/.

the simulations that most closely resemble the empirical distribution are from the population in which the true size of every effect is zero[11].

I call the reader's attention to the simulated literature in which all effects are zero and there is no publication bias (i.e., the bottom left panel of Figures 6 and 7). This simulation illustrates what occurs when pure noise is studied repeatedly with ks and Ns similar to the deception literature. Somewhat contrary to my argument that publication bias is a root cause of the proliferation of false positives, this simulated literature resembles the empirical literature quite closely. That is, even if one assumes the best possible conditions, it remains plausible to obtain the empirical literature while also assuming that there are no real cues to deception.

Thus, publication bias is not an essential component of the Land of Toys *per se*. Rather, what is required is a large number of effects measured with a small number of reported estimates. Or more to the point, what is required is a *lack of precision* in the estimation of a large number of effects. When sample sizes are small, pervasive imprecision can arise due to selective reporting, few attempts at replication, or a combination of the two.

One can also see that, in addition to the null simulated literature, the empirical literature resembles the simulated literatures in which all effects are |0.10|. The reason the empirical literature is so similar to both the |0.10| simulations and the null simulations is that the |0.10| simulations and the null simulations are *similar to each other*. The ks and Ns of the empirical literature (and thus, the simulations) are such that they would not behave very differently if every effect were null or if every effect were |0.10|. This is another symptom of low statistical power in the empirical literature. If there were more power, the null and |0.10| simulations would be more distinct from each other, the estimates would be more accurate, and the distributions would overlap less. We would have a better sense of whether the cue effects were real, because the pattern of the estimates would be different if there were many real effects or none at all. Here, they are so similar that drawing inferences about the existence of effects is extremely prone to error.

Try it yourself. My experience suggests that it is difficult to set parameters for simulations that produce distributions similar to the empirical literature, other than every effect being zero or extremely close to zero, with little to no heterogeneity. For readers interested in testing this for themselves but lacking the computational resources (or the patience) to run more exhaustive Monte Carlo simulations, I have written a Shiny app that allows the user to create less computationally intensive simulated deception cue literatures specifying the following parameters: (1) number of studied cues (up to 200), (2) mean true effect size, (3) standard deviation of true effect sizes, (4) proportion of non-significant results reported, and (5) alpha level. The app can be found here:

[11]The eagle-eyed reader will notice there appear to be two empirical cues whose estimates are outside the area covered by the N-based plot of the null simulations. These two cues are *impressions of verbal immediacy* and *cooperativeness*. Although their effects appear at least somewhat promising, both these cues are heavily influenced by outlier estimates from the same study, namely Horvath, Jayne, & Buckley (1994), whose methods have been criticized for, among other things, insufficiently establishing the ground truth of the rated messages (Vrij, Mann, & Fisher, 2006).

https://rabbitsnore.shinyapps.io/deception_literature_simulator/

I have argued that a critical component of the Land of Toys problem is low power of the studies that make up the literature, but I have only presented simulations under a single set of power conditions, namely conditions similar to those observed in DLMMCC. Informative simulations of deception literatures testing differing samples sizes would be extremely computationally demanding, so I have eschewed such a thorough test. However, to provide less precise simulations, I have written a Shiny app nearly identical to the one above but with the added feature of user-specified sample sizes for all studies in the simulated literature. Trying different sample sizes in the app, it is easy to see that large sample sizes ameliorate the problem substantially, by providing more accurate individual estimates of effect sizes. The app can be found here: https://rabbitsnore.shinyapps.io/deception_literature_simulator_various_sample_size/

Summary. These simulations (and the reader's own experiences with the apps, I trust) indicate that the literature offers extremely weak evidence that truthful and deceptive messages reliably differ. According to visual inspection and quantitative metrics (see Appendix C), the empirical literature is compatible with there being no real cues to deception. The entire literature could have plausibly been sampled from noise.

Conclusions

Theoretical and practical implications. Despite the common interpretation that some behaviors reliably (albeit weakly) correlate with deception, low power, few replications, and publication bias may have produced a literature like the one we have, even if all the true effects we have studied are zero. This does not mean that every cue to deception actually does have an effect of zero; rather, it means that we do not know with confidence that any studied cue does in fact have an effect different from zero. The features of deception cue research – specifically that many cues are measured under conditions of low power, possibly reported selectively, and studied few times – are cause for skepticism in the accuracy not only of estimates in individual studies but also meta-analytic estimates, such as those in DLMMCC.

Individual cue estimates cannot, at present, be trusted. The estimated meta-analytic effect size for any given cue may be quite inaccurate. The simulations indicate that if researchers studied pure noise with the ks and Ns of the deception literature, false positives would proliferate. This does not mean they have; it means they could have. One can think of this situation as analogous to examining a nonsignificant p-value in a study with a small sample. In conventional statistics, we interpret low p-values as suggestive that the null hypothesis may not be true – as if to say, "It would be very strange to get such results if the null hypothesis were true, especially if we consistently get such results." Consider, for example, the implications of $p = .32$ with $N = 40$ (an observed effect of $d = 0.32$). In addition to not being significantly different from 0, these results would also not produce a significant equivalence test (see, e.g., Lakens, 2017). Thus, it would neither support the alternative hypothesis nor the null hypothesis. Its informational value

is low. We are faced with similar results in the aggregated deception literature: We have data that could have plausibly been produced under the null hypothesis, as well as a range of alternative hypotheses. Under such circumstances, we typically refrain from inferring that the null hypothesis is false – nor do we *accept* the null.

There may be many authentic cues to deception, but there is no way, at present, to look at any given deception cue estimate and to be confident that it is not the result of error. It is easy, if not inevitable, to obtain numerous spurious results in the past and current deception cue research paradigm. The informational value of each study in the deception literature has been so low and the possibility for bias and error so high, it is virtually impossible to determine if any given cue's effect is genuine or simply a ghost in the noise. But often when one goes searching for something, one finds it, whether or not it actually exists (Nickerson, 1998). Deception researchers have often (understandably) interpreted individual effects found in DLMMCC and elsewhere as evidence for potentially useful tools to detect deception (e.g., Akehurst, et al., 2017; Evans et al., 2013), but to do so is misguided.

The low trustworthiness of current estimates has implications for future efforts to empirically study cues to deception. Basing hypotheses (and power calculations) on naive readings of reported effect estimates may be fraught with error. That is not to say the situation is hopeless. One might reasonable follow leads that appear promising for theoretical or empirical reasons. As one example, there is a cue right at the edge of the area covered by the null simulations, with a relatively large total sample size. This cue is *details* (i.e., the level of detail or amount of specific information included in a statement), with an estimated effect of $d = 0.30$ (total $N = 883$ from $k = 24$ estimates). In the null simulation with no publication bias, above $N = 800$, the largest effect is about 0.26. Under conditions of publication bias, the largest effect in the null simulations above $N = 800$ is about 0.31. As such, the empirical *details* cue is on the high end of what would be expected if every effect were null[12]. Although one ought to be cautious about making claims about this cue right now, the level of detail in a message may indeed be a promising cue to deception that is worthy of future study.

The Land of Toys problem is not merely academic; there are human and monetary costs to errors of this kind, as when, for example, governmental agencies spend public funds on security policies justified with naive and selective interpretations of the science of deception (see, e.g., Government Accountability Office, 2017; Halsey, 2013) or police and military personnel are trained to make decisions using potentially faulty cues to deception (see, e.g., Kassin & Fong, 1999; Meissner & Kassin, 2002). In the preface to the second edition of his power manual, Cohen (1988) observed that many, if not most, social scientists continued to neglect statistical reforms, and he grimly declared: "They do so at their peril" (p.xv). In deception research, we also do so at the peril of others. Basing practical or policy recommendations on particular cues from the literature – even ones estimated with seemingly large effects – may be unjustified at present. Strong empirical evidence is required to establish the verity of any cue to deception. It is possible that, in

[12]There are several outliers contributing to the size of the estimate for *details* in DLMMCC. This ought to be borne in mind when considering the evidence for this cue.

the many decades of deception research, there has never been such evidence.

Regardless of the trustworthiness of individual cue estimates, however, the established conclusion that deception cues are generally weak is almost assuredly correct. A literature's power to detect effects is partly a function of the true size of the effects being studied. Even with small samples, it is possible to reliably detect large effects. Thus, if strong cues existed, we likely would have found them by now – unless we have been systematically looking in the wrong places. Just as DLMMCC concluded, the observable differences between truthful and deceptive messages are minute. However, this conclusion may be understated. Deception cues may be considerably weaker than their estimates suggest.

Additionally, one would be on shaky ground to claim support for any given theory on the basis of the present data on deception cues. Because we cannot trust individual cue estimates, doubt is thrown onto ostensible patterns of deception cues that might appear to favor some particular theoretical perspective. Patterns estimated with low power do not necessarily replicate patterns in the larger population (Tversky & Kahneman, 1971). Particularly when effects are small in size and power is low, it is easy not only to incorrectly estimate the size of an effect but also to incorrectly estimate whether the effect is positive or negative (Gelman & Carlin, 2014). As such, in the deception literature, there might be many cues that are false positives or overestimates, and there may also be cues for which we currently have the wrong sign. Such potential errors throw theoretical interpretations into serious question.

Is there a villain in this story?

> *We knew many researchers—including ourselves—who readily admitted to dropping dependent variables, conditions, or participants to achieve significance. Everyone knew it was wrong, but they thought it was wrong the way it is wrong to jaywalk. ...[S]imulations revealed that it was wrong the way it is wrong to rob a bank.* - Simmons, Nelson, and Simonsohn (2018, p.255)

> *"Come with us and we'll always be happy," shouted the one hundred and more boys in the wagon, all together.*
> *"And if I go with you, what will my good Fairy say?" asked the Marionette, who was beginning to waver and weaken in his good resolutions.*
> *"Don't worry so much. Only think that we are going to a land where we shall be allowed to make all the racket we like from morning till night."*
> *Pinocchio did not answer, but sighed deeply once—twice—a third time. Finally, he said: "Make room for me. I want to go, too!"* - on the way to the Land of Toys (Collodi, 1883)

Having raked the deception literature over the coals for its methodological shortcomings, it is important to consider the underpinning motives and causes of questionable practices, as they have implications for the development of practical remedies to detect and prevent such practices in the future. Once again, the story of Pinocchio is illustrates the heart of the problem.

Pinocchio did not travel to the Land of Toys with malicious intent; it seemed like a good idea at the time. Moreover, although he knew it was wrong, he did not seem to recognize the high cost of lingering in the Land of Toys. He had been warned; indeed, one of the first warnings he receives from the Talking Cricket is that if he kept behaving as he was, he would become a donkey. He ignored that exhortation (or perhaps did not think about it very hard). He was pulled in two directions – by his motivation to be good and his motivation to enjoy himself. It is easy to understand why he succumbed to temptation: *The benefits were apparent, and the adverse consequences were not.*

So it has been, I believe, in deception research (and in psychological science in general). We have seen that the deception literature exhibits substantial methodological problems: samples sizes are too small to detect plausible effects, results are reported selectively, and the number of significant results appears to have been substantially inflated. This is consistent with the general finding that questionable practices are quite common in science (Agnoli et al., 2017; Fraser et al., 2018; John et al., 2012; Kerr, 1998; Martinson et al., 2005). Why are questionable practices so prevalent? Researchers likely have at least two simultaneous motives for their work: (1) to produce sound science and (2) to personally benefit (e.g., advance their career). These goals often conflict; what advances one often inhibits the other. When people have simultaneous conflicting goals, they tend to either prioritize one over the other or attempt to compromise in such a way that both goals are (perceived to be) at least adequately satisfied (see Kruglanski et al., 2012; Kunda, 1990). Questionable practices are common, I believe, because they are often, but not always, an attempt at compromise between researchers' conflicting goals. Sometimes, researchers simply do not realize they are doing anything questionable, because they are unfamiliar with or misunderstand relevant statistical techniques. Other times, researchers know questionable practices are problematic but engage in them for personal benefit, but they are generally unaware of the *extent* to which the practices are problematic because, again, they inadequately understand statistics.

The personal benefits of questionable practices are obvious. Researchers have incentives to appear productive through publications (Hirsch, 2005; Tijdink et al., 2016a). Publishing positive results is easier than publishing null results (Fanelli, 2012; Giner-Sorolla, 2012). Questionable practices make it easier to obtain positive results (Bakker et al., 2012; Simmons, Nelson, & Simonsohn, 2011). Therefore, researchers have incentives to engage in questionable practices. These incentives induce researchers to cut corners in their work (Tijdink et al., 2014) and make researchers more likely to report positive results (Fanelli, 2010), with inflated effect sizes (Fanelli et al., 2017). Vazire (2015) describes an evocative anecdotal experience of being tempted by questionable practices *in situ*, and it is striking how obvious the benefits would be and how one could justify the decision to oneself and others (and the justifications seem so compelling in the moment!). For me and likely many others, this is a familiar experience.

Social norms may also influence the commission of questionable practices (Rajah-Kanagasabai & Roberts, 2015). That is, if researchers operate in an environment in which poor practices are perceived as normal and common, they may be more willing to engage in them. Additionally, there is general social consensus among

researchers concerning the relative wrongness of different questionable practices, such that if one admits to one bad practice, one is much more likely to admit to less severe ones as well (Agnoli et al., 2017; John et al., 2012). There is, however, substantial variation in how far individual researchers are willing to go. Providing incentives for truthful disclosure increases rates of admissions of engaging in questionable practices (John et al., 2012), suggesting that researchers know there is a reason to conceal them. Moreover, researchers generally perceive questionable practices to be indefensible, except when they themselves have committed them, in which case they are much more likely to say they are "possibly defensible" (Agnoli et al., 2017; John et al., 2012). The incentives for questionable practices are many, but researchers are not relentlessly opportunistic. Occasionally, researchers may overtly and knowingly engage in wrongdoing. However, the data suggest that generally researchers try to act within the constraints of defensibility, sometimes striking compromises between the goals of scientific soundness and personal benefit, attempting to satisfy both objectives.

Even when one is trying to act in a scientifically rigorous manner, it is possible to err due to lack of knowledge. One's perception of what practices are defensible is likely influenced by one's understanding of the consequences of questionable practices. But surveys consistently find that researchers generally have inadequate or incorrect understandings of statistical concepts relevant to their work (e.g., Bakker et al., 2016; Gigerenzer, 2004; Greenland et al., 2016; Tversky & Kahneman, 1971). How does one square this with the fact that, for decades, there have been clear, consistent, and repeated warnings about low power (e.g., Cohen, 1962, 1969, 1988; Lane & Dunlap, 1978) and selective reporting (e.g., Greenwald, 1975; Rosenthal, 1975; Sterling, 1959)? Wouldn't well-intentioned scientists leap at the opportunity to improve their methods? Despite the availability of knowledge on how to improve practice, it remains remarkably easy for researchers to fail to appreciate the consequences of their decisions as they are making them, particularly when poor practices are not met with resistance (e.g., critique during peer review).

In their classic paper on the "law of small numbers," Tversky and Kahneman (1971) describe overestimation of the reliability of statistical trends (and thus overestimation of power) as a pervasive foible of human reasoning. They note that people commit a "multitude of sins against the logic of statistical inference *in good faith*" (p.110, emphasis added). If researchers insufficiently understand the statistics they are using, and instead uncritically follow conventions established by others who have a similarly poor understanding (Gigerenzer, 2004; Sedlmeier & Gigerenzer, 1989), they may fail to appreciate the gravity of the poor practices. That is, the deleterious consequences of questionable practices may be downplayed in part because they appear to work just fine, and they do not have the requisite knowledge to realize they are not actually working properly (see Nelson, Simmons, & Simonsohn, 2018). For example, researchers may conduct unplanned hypothesis tests that inflate the false positive rate as they attempt to understand their data and do not fully realize just how wrong their decisions are (Gelman & Loken, 2013). Thus, researchers may be more willing to engage in questionable practices, as they are less aware of the threat to their motivation to produce sound science (Agnoli et al., 2017; John et al., 2012). That is, they might not realize the wrongness of

questionable practices at all, or if researchers do know they are bad, they might not seem bad enough not to do them.

Along these lines, it is possible that questionable practices themselves influence the feedback scientists receive from their own work (Nelson, Simmons, and Simonsohn, 2018), such that it seems one can regularly succeed, for example, with extremely small samples. Evidence reviewed above demonstrated a higher-than-expected rate of significant results in the deception literature, and published articles with no significant cue effects often reporting significant effects when analyzing subsets of their data or collapsing across conditions. To individual deception researchers, it may well appear that their work has been successful. Pinocchio's invitation to the Land of Toys was a ringing endorsement of its many pleasures, with the costs downplayed and dismissed. Similarly, the process by which problematic research decisions are made may in itself conceal the hazards therewith.

As I see it, there is no villain in the story of deception research: Our problems are self-generated, but our poor decisions have also been understandable. That does not change the fact our research practices have not lived up to proper standards. Pinocchio was still transformed into a donkey, even though he realized he should have heeded the warnings he had earlier received. Perhaps it is not entirely our fault, but it is certainly our responsibility to fix it. No matter how we wish to litigate the blameworthiness of past decisions, we will have to deal with the consequences.

We must change the way we do things. From a certain point of view, I have presented nothing new here. The causes of the Land of Toys problem – that is, low power, selective reporting, and a lack of replications – have been known to be deleterious for a long time (Cohen, 1962, 1988, 1994; Easterbrook et al., 1991; Fisher, 1926; Lane & Dunlap, 1978; Meehl, 1990; Open Science Collaboration, 2015; Rosenthal, 1979; S. Schmidt, 2009). It is not that we have not known these methodological issues were problematic; it is that we have not paid adequate attention to them (Nelson, Simmons, & Simonsohn, 2018).

Solutions to the Land of Toys problem are, like the Talking Cricket and Turquoise Fairy's advice, thankfully readily available. The methodological reforms necessary to avert (or at least substantially reduce) the Land of Toys problem are not radical or novel: (1) Researchers must more fully disclose their methods, analyses, and data, and (2) researchers must improve the statistical power of their work. These reforms are demanding but not complex. Fortunately, we also have even more convenient tools at our disposal to mitigate these problems, compared to decades ago. That said, implementing the solutions will likely require substantial changes to the way deception research has been typically conducted.

The first recommendation calls for less selective reporting. Reducing publication bias is challenging, in part because at least some of this bias is out of the hands of individual researchers. Fortunately, authors that can ameliorate the problems of publication bias and selective reporting in a variety of ways. Preprint repositories (e.g., OSF Preprints, PsyArXiv) and blogging make it possible to more rapidly and widely disseminate research findings, especially those that might otherwise be suppressed. Open data practices can also greatly mitigate selective reporting problems (Miguel et al., 2014). If, for example, space restrictions prevent authors from exhaustively reporting extensive results, making data openly available allows

other researchers to explore for themselves.

Preregistration of coding and analyses would also mitigate this issue by reducing researchers' ability to select cues that might favor their hypotheses and their ability to change their hypotheses to suit the data (Kerr, 1998; Nosek et al., 2018). If they are not already, deception researchers ought to become familiar with the myriad ways they can inadvertently arrive at mistaken conclusions. Here, theory can be helpful as well. Specifically, it can provide guidance about what cues might be worth studying and can help shape meaningful interpretations of results. For example, if joined with preregistration, well-specified theory can help curtail problematic *ad hoc* explanations for results. In short, increasing transparency in research will substantially improve the deception literature.

The second recommendation calls for increases in statistical power (and thus, precision). We can see from the simulations that a literature of false positives can grow even without publication bias, but such bias exacerbates the problem by reducing the likelihood that spurious results will be corrected in the long run. However, for there to be a chance of correcting a false positive in the long run, there must be a long run in the first place. That is, deception researchers must repeatedly study and report cues, rather than relying on small numbers of estimates, especially from small samples.

Raising statistical power is largely a matter of increasing sample sizes. Deception research has been outrageously underpowered. This problem cannot be over-emphasized. Let us assume – admittedly unsafely – a typical effect of 0.10 (the median in DLMMCC). For .80 power to detect an effect that size, we would need $N = 3{,}142$. Recall that the average sample size in DLMMCC was approximately $N = 41$. This is alarmingly small, given the probable size of the effects. There is no escaping mathematical reality. If we want to detect and estimate small effects, we are going to need to face this challenge. We will need *much* larger samples to verify the existence or nonexistence of cues to deception.

Under some circumstances, within-subjects designs, in which multiple messages are obtained from each participant, can help raise power by increasing the number of sampled observations (see Judd, Westfall, & Kenny, 2017). Unfortunately, even with somewhat more powerful within-subjects designs, it may be unlikely that any single research group could consistently acquire the resources necessary for adequate power to study very small effects like the ones that may be present in deception. The same problem obtains for focused efforts to directly replicate results. For this reason, "many labs" collaborations, in which numerous research teams pool their resources, may be a practical solution to these challenges (e.g., Klein et al., 2014). Thankfully, such large scale collaborations are not unheard of in deception research (e.g., Global Deception Research Team, 2006). Their implementation, however, must become more commonplace if we are to leave the Land of Toys.

Most of these recommendations are old. Psychological scientists have generally been resistant to statistical and methodological reforms (Rossi, 1990; Sharpe, 2013), and I have discussed above the ways in which poorer practices offer personal benefits. The consistent patterns of small samples and selective reporting and the unusually high rate of significant effects demonstrate that deception researchers have not, in spite of recommendations, voluntarily implemented reforms. Thus, to

change the face of research practice, rather than trusting researchers to voluntarily adopt reforms, it is likely that the general standard of evidence must be raised and the incentives for good research practices must be changed with policy (see Nosek, Spies, & Motyl, 2012; Sedlmeier & Gigerenzer, 1989).

Given the ease with which one can obtain spurious results, editors and reviewers may, for example, need to require more rigorous application of the practices described above (e.g., preregistration, accumulation of substantial evidence) in order to publish strong claims about cues to deception (see Giner-Sorolla, 2012; Vazire, 2018a). Many of us may find such demands unpalatable (Vazire, 2018b), but substantive changes in methodology are unlikely to occur if individual researchers do not face consequences for suboptimal methods (Smaldino & McElreath, 2016). To the extent that evidence is required to justify such demands on researchers, simulations and reviews of the literature like the one presented here might provide such policy justifications.

The rabbits have arrived

To the extent that these recommendations entail practices that differ from the typical manner in which deception research has been conducted, they may be challenging to adopt. Their potential difficulty, of course, makes them no less necessary. When he was dying from injuries, Pinocchio refused to drink the foul-tasting medicine offered by the Fairy with Turquoise Hair. He declared that he would rather die than endure the wretched taste of the medicine that would save his life. Then a procession of four black rabbits bearing a coffin on their shoulders arrived and informed him, "We have come for you." It was only then, faced with a clear signal of his impending doom, that he took the medicine, which promptly healed him just as the Fairy had told him it would.

As I have explained above, I suspect that deception researchers, like Pinocchio, have not taken the bitter medicine of methodological reform in large part because the negative consequences of present practices have not been apparent. My hope is that what I have presented here will serve as a clear indication that the current way of doing things cannot be sustained. We must hold ourselves to higher standards and demand stronger evidence for claims of cues to deception. Effective reforms will be onerous. The medicine is bitter. But the alternative is worse.

Open Practices

Code to reproduce all simulations and figures, code for the Shiny apps, and data produced by the simulations and extracted from DLMMCC are available on the Open Science Framework (https://osf.io/cf5vs/). As of writing, I have not received permission to publicly post the original data for DLMMCC, so I have not. Should I receive permission in the future, the data and code to reproduce relevant analyses will be available on OSF (https://osf.io/mfq6u/).

References

Akehurst, L., Easton, S., Fuller, E., Drane, G., Kuzmin, K., & Litchfield, S. (2017). An evaluation of a new tool to aid judgements of credibility in the medico-legal setting. *Legal and Criminological Psychology, 22*, 22-46.

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*(3), e0172792.

Amado, B. G., Arce, R., Farina, F., & Vilarino, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology, 16*, 201-210.

Anolli, L., & Ciceri, R. (1997). The voice of deception: Vocal strategies of naive and able liars. *Journal of Nonverbal Behavior, 21*, 259-284.

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

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*, 1069-1077.

Bond, C. F. Jr., Levine, T. R., & Hartwig, M. (2014). New Findings in Non-Verbal Lie Detection. In P.A. Granhag, A. Vrij, & B. Verschuere (eds), *Detecting deception: Current challenges and cognitive approaches* (pp.37-58). New York: Wiley.

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.

Burns, J. A., & Kintz, B. L. (1976). Eye contact while lying during an interview. *Bulletin of the Psychonomic Society, 7*, 87-89.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2018, July 23). Correcting for bias in psychology: A comparison of meta-analytic methods. https://doi.org/10.31234/osf.io/9h3nu

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*, 997-1003.

Cohen, J. (1990). Things I Have Learned (So Far). *American Psychologist, 45*, 1304-1312.

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

Cohen. J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Collodi, C. (1883). *Adventures of Pinocchio.* Available at http://www.gutenberg.org/ebooks/500

Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25*, 7-29.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230-232.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118.

DePaulo, B. M., Rosenthal, R., Green, C. R., & Rosenkrantz, J. (1982). Diagnosing deceptive and mixed messages from verbal and nonverbal cues. *Journal of Experimental Social Psychology, 18*, 433-446.

Duval, S. J. & Tweedie, R. L. (2000). Trim and fill. A simple funnel-plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-463.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet, 337*, 867-872.

Ekman, P., Friesen, W. V., & Simons, R. C. (1985). Is the Startle Reaction an Emotion? *Journal of Personality and Social Psychology, 49*, 1416-1426.

Evans, J. R., Michael, S. W., Meissner, C. A., & Brandon, S. E. (2013). Validating a new assessment method for deception detection: Introducing a Psychologically Based Credibility Assessment Tool. *Journal of Applied Research in Memory and Cognition, 2*, 33-41.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS one, 5*(4), e10271.

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, 201618569.

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics, 33*, 615-630.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33*, 503-513.

Frank, M. G., Menasco, M. A., & O'Sullivan, M. (2008). Human behavior and deception detection. *Wiley Handbook of Science and Technology for Homeland Security*. Wiley.

Fraser, H., Parker, T. H., Nakagawa, S., Barnett, A., & Fidler, F. (2018, March 21). Questionable Research Practices in Ecology and Evolution. https://doi.org/10.1371/journal.pone.0200303

Gelman, A., & Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time.* Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562-571.

Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology, 37*, 60-74.

Government Accountability Office (2017). *TSA Does Not Have Valid Evidence Supporting Most of the Revised Behavioral Indicators Used in Its Behavior Detection Activities.* Retrieved from https://www.gao.gov/products/GAO-17-608R

Greene, J. O., O'Hair, H. D., Cody, M. J., & Yen, C. (1985). Planning and control of behavior during deception. *Human Communication Research, 11*, 335-364.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology, 31*, 337-350.

Greenwald, A. G. (1975). Consequences of Prejudice Against the Null Hypothesis. *Psychological Bulletin, 82*, 1-20.

Halsey, A. III (November 13, 2013). GAO says there is no evidence that a TSA program to spot terrorists is effective. *Washington Post.* Retrieved from https://www.washingtonpost.com/local/trafficandcommuting/gao-says-there-is- no-evidence-that-a-tsa-program-to-spot-terrorists-is-effective/2013/11/13/fca999a0-4c93-11e3-be6b-d3d28122e6d4_story.html

Hartwig, M., & Bond Jr, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology, 28*, 661-676.

Hartwig, M., & Bond, C. F. Jr, (2011). Why Do Lie-Catchers Fail? A Lens Model Meta-Analysis of Human Lie Judgments. *Psychological Bulletin, 137*, 643-659.

Heilveil, I. (1976). Deception and pupil size. *Journal of Clinical Psychology, 32*, 675-676.

Heilveil, I., & Muehleman, T. J. (1981). Nonverbal Clues To Deception In A Psychotherapy Analogue. *Psychotherapy: Theory, Research & Practice, 18*, 329-335.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences, 102*, 16569-16572.

Horvath, F., Jayne, B., & Buckley, J. (1994). Differentiation of truthful and deceptive criminal suspects in behavior analysis interviews. *Journal of Forensic Science, 39*, 793-807.

IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of clinical epidemiology, 68*, 860-869.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124.

Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal, 176*, 1091-1096.

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.

Johnston, S., Candelier, A., Powers-Green, D., & Rahmani, S. (2014). Attributes of truthful versus deceitful statements in the evaluation of accused child molesters. *Sage Open, 4*(3), 2158244014548849.

Judd, C.M., Westfall, J., & Kenny, D.A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601-625

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior, 23*, 499-516.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142-152.

Kruglanski, A. W., Belanger, J. J., Chen, X., Kopetz, C., Pierro, A., & Mannetti, L. (2012). The Energetics of Motivated Cognition: A Force-Field Analysis. *Psychological Review, 119*, 1-20.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*, 355-362.

Lane, D. M.,& Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31*, 107-112.

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333*, 597-600.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). *Scientists behaving badly. Nature, 435*, 737-738.

Mazzola, J. J., & Deuling, J. K. (2013). Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles. *Industrial and Organizational Psychology, 6*, 279-284.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26*, 469-480.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... & Laitin, D. (2014). Promoting transparency in social science research. *Science, 343*(6166), 30-31.

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable.* Preprint available at https://raw.githubusercontent.com/richarddmorey/psychology_resolution/master/paper/response.pdf

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*, 511-534.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*, 175-220.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences,*

201708274.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Phillips, C. V. (2004). Publication bias in situ. *BMC Medical Research Methodology, 4:20*.

RabbitSnore (2018). An attempt to reproduce a meta-analytic review of Criterion-Based Content Analysis. *Rabbit Tracks*. Retrieved from https://www.rabbitsnore.com/2018/06/an-attempt-to-reproduce-meta-analytic.html

Rajah-Kanagasabai, C. J. & Roberts, L. D. (2015). Predicting self-reported research misconduct and questionable research practices in university students using an augmented Theory of Planned Behavior. Frontiers in psychology, 6, 535.

Richard, F. D., Bond, C. F. Jr., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology, 7*, 331-363.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin, 86*, 638-641.

Rossi, J. S. (1990). Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology, 58*, 646-656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin, 105*, 309-316.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90-100.

Schmidt, F. L. (1992). What Do Data Really Mean?: Research Findings, Meta-analysis, and Cumulative Knowledge in Psychology. *American Psychologist, 47*, 1173-1181.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods, 18*, 572-582.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*, 255-259.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January). *Life after p-hacking.* Paper presented at the 14th Annual Meeting of the Society for Personality and Social Psychology, New Orleans, LA. doi:10.2139/ssrn.2205186

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.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*(9), 160384.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology, 53*, 1119-1129.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association, 54*, 30-34.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*, 670-688.

Tijdink, J. K., Schipper, K., Bouter, L. M., Pont, P. M., De Jonge, J., & Smulders, Y. M. (2016a). How do scientists perceive the current publication culture? A qualitative focus group interview study among Dutch biomedical researchers. *BMJ open, 6*(2), e008681.

Tijdink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics, 9*, 64-71.

Tversky, A., & Kahneman, D. (1971). *Belief in the law of small numbers. Psychological Bulletin, 76*, 105-110.

Vazire, S. (2018a). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*, 411-417.

Vazire, S. (2018b). bitter carrots. *sometimes i'm wrong.* Retrieved from https://sometimesimwrong.typepad.com/wrong/2018/05/bitter-carrots.html

Vazire, S. (2015). this is what p-hacking looks like. *sometimes i'm wrong.* Retrieved from http://sometimesimwrong.typepad.com/wrong/2015/02/this-is-\what-p-hacking-looks-like.html

Vrij, A. (2015). Verbal Lie Detection tools: Statement validity analysis, reality monitoring and scientific content analysis. In P.A. Granhag, A. Vrij, B. Verschuere (eds.), *Detecting Deception: Current Challenges and Cognitive Approaches* (pp. 3-35). John Wiley & Sons.

Vrij, A. (2006). Nonverbal communication and deception. In V. Manusov, & M. Patterson (Eds.), *The Sage handbook of nonverbal communication* (pp. 341-359). Thousand Oaks, California: Sage.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and criminological psychology, 9*, 159-181.

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition, 1*, 110-117.

Vrij, A., Mann, S., & Fisher, R. P. (2006). An empirical test of the behaviour analysis interview. *Law and human behavior, 30*, 329-345.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274-290.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al.(2009). *Perspectives on Psychological Science, 4*, 294-298.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and Nonverbal Communication of Deception. *Advances in experimental social psychology, 14*, 1-59.

Appendix A

It might be desirable to estimate the magnitude of publication bias in the deception literature, as such an estimate would permit an empirical evaluation of the extent to which the effects reported in the literature may have been influenced by bias. Can we estimate publication bias using the data from DLMMCC? Numerous conceptual and statistical difficulties stand in the way of this task in the deception cue literature. The peculiar methodology of deception cue research lead me to conclude that attempting to estimate and correct for publication bias in this literature is at best extremely challenging and at worst futile.

Meta-meta-analytic approach

To assess publication bias, one might construe the deception cue literature as a body of data examining a single phenomenon – as a meta-meta-analytic problem. That is, one could meta-analyze the meta-analytic estimates ($k = 158$) and apply techniques to address publication bias on the data reported in DLMMCC. However, this approach is misguided.

First, there is a conceptual issue. It unclear what publication bias (or lack thereof) at this level of analysis means. It is clear what it does *not* mean, however. Publication bias at the meta-meta-analytic level does not directly reflect bias in the reporting of individual studies or effect size estimates for cues. Rather, at this level of analysis, publication bias may reflect systematic exclusion of entire cues from the literature. It is possible (if not likely) that some cues have simply never been reported. However, it is not clear that standard publication bias assessment and correction techniques would be appropriate to address this possibility.

Second, there is a methodological issue. In DLMMCC, the signs of the effect sizes are sometimes arbitrary. They were coded such that positive effects indicated an increase in the given cue in deceptive messages, compared to truthful messages. For some cues that are measured in the frequency or duration of a given behavior, such as smiling or hand movements, the sign of the effect takes on an intuitive meaning. For other cues that deal with quantified impressions, such as impressions of nervousness, immediacy, or cooperativeness, the sign is an artifact of the labeling of the variable. If the labels were inverted (e.g., "calmness," "hesitation," and "uncooperativeness"), the sign of the effect would flip, with no loss of meaning. This poses an obvious problem for publication bias techniques that are influenced by the signs of effect estimates (e.g., trim and fill, PET-PEESE, selection models; see, e.g., Carter et al., 2018), and it renders meta-meta-analytic statistics untrustworthy.

To address this issue, one might consider taking the absolute value of each estimate and conducting analyses on the resulting data, all of which would have positive signs. This would solve one problem and introduce others. Specifically, this approach implicitly assumes that the empirical research has obtained the correct sign for each cue (correctness, that is, with respect to however the cue has been labeled). For estimates close to 0 (of which there are many), which could easily have the wrong sign, this assumption is questionable (see Gelman & Carlin, 2014). If sign errors have occurred, absolute values will overestimate effects.

Taking the absolute value of every effect would also have the result of essentially folding the funnel plot in half (i.e., everything on the left side would be reflected onto the right side). The artificial asymmetry of the distribution would render many estimates of publication bias meaningless (again, not that they are necessarily informative at this level of analysis anyway).

Meta-meta-analyses of the DLMMCC cue classifications. A possibly more appropriate approach is to use the categories into which DLMMCC sorted deception cues (i.e., forthcoming, compelling, positive, tense, and ordinary imperfections). Within each category, the signs of the effects were not arbitrary, as they corresponded to the increase or decrease, in deceptive behavior relative to truthful behavior, of the construct identified by the category. Therefore, taking this approach would avoid the sign problem, but it will also not solve the conceptual problem that publication bias at the meta-meta-analytic level does not effectively capture bias at the level of the individual study. However, by avoiding the arbitrary sign problem, this approach can provide a potentially useful summary of the size of effects of deception cues better than one big meta-meta-analysis. Results of such a summary are reported in Appendix B.

Summary. In short, a meta-meta-analytic approach to estimating publication bias in the deception literature is fraught with difficulties.

Additional note. The astute reader may wonder if the matter of arbitrary signs threatens the validity of inferences drawn from the simulated deception cue literatures. Thankfully, it does not. The simulated literatures generated to assess the Land of Toys problem are all necessarily symmetrical; they have equal amounts of positive and negative effects, and the signs of the simulated effects are arbitrary, as many are in DLMMCC. For this reason, if one considers it desirable to take the absolute values of the effects in DLMMCC for analysis, one can also take the absolute values of the simulated effects for comparison. The metrics of similarity presented in Appendix C are invariant to the signs of the effects.

My own preference is to examine the raw effect estimates, with original signs intact, because I find it more intuitive to examine the symmetrical plots, which represent the way the accumulating cue estimates deviate on either side of 0. However, for readers who prefer otherwise, alternate versions of Figure 6 and 7 which plot absolute values are available here: https://osf.io/gfhqe/

Publication bias in individual cue estimates

Alternatively, one might consider publication bias at the level of each cue. That is, one could apply statistical techniques to address bias individually, for every one of the 158 cues in DLMMCC. This circumvents the conceptual problems described above, and it is conceptually in line with the type of publication bias relevant to the issues addressed in this paper. However, it runs up against another problem. Specifically, most deception cues have been reported only a small number of times, greatly reducing the effectiveness of publication bias metrics (see, e.g., Ioannidis & Trikalinos, 2007; Lau et al., 2006; Sterne, Gavaghan, & Egger, 2000). Of the 158 reported cues, 42 were reported only once, rendering publication bias estimates impossible for those. That leaves 116 potentially viable cues. The majority of these potentially viable cues (65) were only reported between 2 to 5 times – offering little

power to detect bias. Only 26 cues were reported 10 or more times, and only 9 cues were reported 20 or more times. The situation is worsened when one considers that DLMMCC were unable to extract exact effect sizes in many instances. Only 18 of the cues could be precisely extracted 10 or more times, and 71 could only be precisely extracted between 2 and 5 times. As such, the statistical power of publication bias tests for each cue would be woefully low.

Perhaps estimating publication bias where possible (i.e., for higher-k cues) seems as if it would be informative. However, this approach would in fact be highly problematic. As noted above, the distribution of effect size estimates seems to center on or close to 0. For cues reported at least $k = 10$ times, the average effect size is similarly close to 0, but there is much less variation ($SD = 0.132$ vs. $SD = 0.265$ for the full distribution; and see Figure 5 for a visualization). The ambition of these bias assessment procedures would be to obtain an estimate of bias that could be generalized to those cues for which it is not appropriate to apply such techniques. But the ability to employ techniques to assess publication bias (deriving from the number of reported studies) is confounded with the size and heterogeneity of estimates, such that we have the least ability to accurately assess publication bias for the cues for which we would be most concerned about publication bias (i.e., those with estimated effect sizes that substantially deviate from 0).

As an additional note, consider that the the results of the large-scale deception literature simulation (presented in Figures 6 and 7 and in Appendix C) indicate that even under conditions of low or no publication bias, it is still highly plausible to have obtained the empirical literature when all cues have effects of zero. As such, with regard to the question of the evidential value of the extant literature, attempting to approximate publication bias is something of a moot point.

Appendix B

DLMMCC classified deception cues into five categories. These categories were established to assess whether liars, compared to truth-tellers, (1) were less forthcoming, (2) told less compelling tales, (3) were less positive, (4) were more tense, and (5) told stories with fewer ordinary imperfections. Table 2 displays some descriptive information about the literature in these five categories. Within each of the five categories, the signs of the cue effects were not arbitrary (see Appendix A); rather, the signs of the effects indicated an increase (among liars relative to truth-tellers) of the construct in question (e.g., positivity, tension). Thus, one could attempt to summarize the deception cue literature writ large as five meta-meta-analyses, one for each type of cue in DLMMCC, each analysis providing an estimate of the strength and direction of the typical deception cue in each category.

Table 1: *Descriptive information for deception cue categories from DLMMCC*

Cue type	Total cues	Mean N (SD)	Total N	Mean k (SD)	Total k
Forthcoming	14	454.29 (547.89)	6,360	11.93 (14.45)	167
Compelling	65	264.51 (297.88)	17,193	6.32 (7.29)	411
Positive	18	330.00 (323.67)	5,940	6.39 (6.48)	115
Tense	12	444.00 (275.29)	5,328	9.83 (6.10)	118
Ordinary imperfections	19	140.68 (67.56)	2,673	3.79 (1.96)	72

Note: Total cues = number of cues in the category. Mean N = average sample size per cue (i.e., sum of all studies). Total N = total sample size of all cues. Mean k = average number of studies per cue. Total k = total number of studies for all cues.

Table 2: *(Meta-)meta-meta-analytic summaries of deception cue categories from DLMMCC*

Cue type	Estimate	95% CI	z	p-value
Forthcoming	-0.078	[-0.172, 0.015]	1.648	0.099
Compelling	-0.002	[-0.052, 0.047]	0.092	0.926
Positive	-0.070	[-0.164, 0.023]	1.469	0.142
Tense	0.099	[0.004, 0.194]	2.046	0.041
Ordinary imperfections	-0.145	[-0.270, -0.019]	2.260	0.024
Overall (absolute values)	|0.064|	[0.116, 0.012]	2.423	0.015

Note: Resources and code to reproduce these analyses are available at https://osf.io/mfq6u/.

Table 2 presents the results of just such a meta-meta-analytic synthesis for each of the five categories of cues. Additionally, I conducted a meta-meta-meta-analysis of the absolute value of the estimates from the five meta-meta-analyses, which can provide an estimate of the overall strength of deception cues reported in DLMMCC. Each estimate was produced using a random effects model. As can be seen in the table, the effect of the typical deception cue in each category (and overall) is hardly impressive – all of them negligible by conventional benchmarks. These results suggest that the effects of deception cues are generally puny.

Note that these estimates are taken at face value from DLMMCC and include no attempt to correct for bias. Given the problems noted in Appendix A, conventional estimates or corrections of publication bias at this level of analysis is likely to be uninformative. However, it is worth noting that the absolute values of the meta-meta-analytic estimates are nearly perfectly correlated with their standard errors, $r = .98$, such that the more precise estimates of the five are closer to zero and the less precise estimates are further away. This does not directly assess the presence of publication bias or whether the effects are spurious, but it does not inspire confidence in the authenticity of these cues.

Appendix C

Table 3: *Quantitative metrics of similarity between empirical and simulated literatures*

***k*-based approach**					
Proportion of n.s. reports included	Effect population				
	0	|0.10|	|0.25|	Normal (0, .25)	Normal (0, .50)
0.10	0.475	0.493	0.410	-0.432	-0.593
0.33	**0.669**	0.593	0.575	-0.101	-0.404
0.50	0.662	0.629	0.586	0.086	-0.248
0.66	0.652	0.654	0.609	0.286	-0.032
0.80	**0.694**	0.657	0.637	0.555	0.328
1.00	**0.697**	0.662	0.632	0.650	0.615
***N*-based approach**					
Proportion of n.s. reports included	Effect population				
	0	|0.10|	|0.25|	Normal (0, .25)	Normal (0, .50)
0.10	0.589	0.561	0.563	-0.572	-0.738
0.33	0.767	0.721	0.760	-0.233	-0.505
0.50	**0.837**	0.788	0.752	0.312	-0.290
0.66	**0.833**	0.798	0.754	0.680	0.198
0.80	0.830	0.799	0.770	0.737	0.505
1.00	**0.831**	0.797	0.762	0.760	0.756

Note: Each coefficient represents strength of matching between the given simulated literature and the empirical literature. Upper part of table presents metrics for the k-based approach. Lower part presents metrics for N-based approach. See text for an explanation of how the metrics were calculated. Three strongest coefficients in each approach, representing the best matches with the empirical literature, are displayed in bold. Resources and code to reproduce these results can be found at https://osf.io/gfhqe/

Although visual inspection is informative, the reader may find it desirable to quantitatively compare the shapes and densities of the simulated distributions and the empirical distribution. To offer a numerical metric of similarity, I took two approaches: one based on the number of studies per cue (k-based) and one based on the total sample size per cue (N-based). For the k-based metric, I split both the effect size data from DLMMCC into sets of cues with a shared number of reports (a total of 25 unique ks). For each level of k, I calculated a measure of deviations from zero, equal to $\sqrt{\frac{SS}{N}}$, where SS is the sum of squared deviations of each effect size from 0 and N is the number of effect sizes. This is in essence a standard deviation from 0 rather than the mean[13], which provides a measure of the spread

[13]That being said, the grand mean of all simulated distributions is 0, and the grand mean of the empirical distribution is very close to 0 anyway, so using standard deviations from the grand mean would produce almost identical results.

of the effect sizes around zero. Calculating a deviation score for each unique k creates a vector of values which characterize the shape of the distribution.

From each of the 30 simulated literatures, I created matching sets of simulated data with the same ks as the empirical data. I then calculated a conventional Pearson correlation coefficient for each of the 30 sets of deviation scores for the simulated literatures with the empirical literature. I then weighted the each coefficient by the number of valid deviation scores obtained from the simulated literature, such that the coefficients were penalized if the simulated literature had a lower maximum k than the empirical literature. Thus, each of the 30 simulations received a single coefficient that attempts to quantify the degree of matching between the empirical literature and that simulation. The coefficients are interpreted such that stronger positive coefficients indicate greater matching between the shape of the empirical and simulated literatures. For the N-based metric, I used the same procedure, but instead of calculating deviation scores for each unique k, I split the data into intervals of total sample size (breaks every 100 participants in the sample, for a total of 14 intervals). I then calculated scores for the effect sizes for cues in each interval and correlated the empirical and simulated scores. Code to reproduce these calculations is available at https://osf.io/gfhqe/.

Table 3 presents the results of these calculations. As in Figures 6 and 7, each row represents a publication bias condition (i.e., the proportion of nonsignificant effects included in the literature), and each column represents a population of effect sizes. As can be seen, the k-based and N-based correlation metrics produced similar (though not identical) results: The best matching simulated literatures by both metrics are those in which all effects are $d = 0$, though the k- and N-based approaches disagree somewhat regarding the best matching publication bias conditions. The correlation metrics also indicate high matching between the $|0.10|$ simulations and the empirical literature. As is noted in the main text, this is unsurprising, given that the null simulations and $|0.10|$ simulations produce such similar results. That is, the empirical literature could not closely match one without matching the other. This is the result of relatively low statistical precision.

One should approach these metrics with caution. Although I believe they represent a reasonable approach to quantifying the similarity in the distributions, they have not been thoroughly validated. One can see that even the simulated literatures that produced results highly visually discrepant with the empirical literature (e.g., the normally distributed effects) received relatively strong coefficients under some conditions. Thus, this approach may be less sensitive to differences in variability than would be desirable. There may be superior ways of assessing the correspondence between the empirical and simulated literatures. I encourage interested others to explore alternatives using the available data.