Research practices that can prevent an inflation of false-positive rates

Article

Accepted Version

Running head: PREVENTION OF FALSE-POSITIVES

Research Practices that can Prevent an Inflation of False-positive Rates

Kou Murayama
University of California, Los Angeles

Reinhard Pekrun
University of Munich

Klaus Fiedler
University of Heidelberg

Published in *Personality and Social Psychology Review*

Correspondence address:
Kou Murayama

Department of Psychology
University of Reading
Earley Gate, Whiteknights
Reading RG6 6AL, UK
Tel: +44 (0)118 378 5558

Abstract

Recent studies have indicated that research practices in psychology may be susceptible to factors that increase false-positive rates, raising concerns about the possible prevalence of false-positive findings. The present article discusses several practices that may run counter to the inflation of false-positive rates. Taking these practices into account would lead to a more balanced view on the false-positive issue. Specifically, we argue that an inflation of false-positive rates would diminish, sometimes to a substantial degree, when researchers 1) have explicit a priori theoretical hypotheses, 2) include multiple replication studies in a single paper, and 3) collect additional data based on observed results. We report findings from simulation studies and statistical evidence that support these arguments. Being aware of these preventive factors allows researchers not to overestimate the pervasiveness of false-positives in psychology, and to gauge the susceptibility of a paper to possible false positives in practical and fair ways.

Keywords: false-positives, Type-1 error, optimal stopping, replication

Research Practices that can Prevent an Inflation of False-positive Rates

There has been no other time when research practices in social psychology (or psychology in general) have been paid such considerable attention and scrutinized to such a great extent (Lilienfeld, 2012; Pashler & Wagenmakers, 2012; Spellman, 2012a). Although concerns about current research practices take myriad forms, one fundamental concern behind many current debates pertains to *false positives*—the incorrect rejection of a null hypothesis. Several critical articles published recently have indicated that research practices in psychology (and other related areas) are susceptible to a variety of factors that increase false-positive rates, which make it difficult to draw valid and reliable scientific conclusions (Bakker & Wicherts, 2011; Fiedler, 2011; Garcia-Perez, 2012; Ioannidis, 2005; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkielman, & Pashler, 2009; Young, Ioannidis, & Al-Ubaydli, 2008). More recently, some special issues devoted to this topic were also published (e.g., Pashler & Wagenmakers, 2012; Spellman, 2012b). We concur with the basic argument advanced in these publications; researchers sometimes, either intentionally or unintentionally, take advantage of factors that can promote false positives. The seriousness of false-positives cannot be over-emphasized—such incorrect findings not only hinder any valid understanding of human nature but also can waste vast amounts of resources for those who believe in false-positive findings.

However, what is missing in the current debate is the explicit recognition of factors in conventional research practice in social psychology that may go *against* the inflation of false-positive rates. Indeed, as shown later, such factors sometimes make a substantial contribution to inhibiting the inflation of Type-1 error rates, making false-positive findings less likely. The purpose of the current article is to discuss such factors, with the aim of promoting a more balanced view on the false-positive issue in social psychology, and of drawing researchers' attention toward ways of fairly gauging the susceptibility of reported research to false positives.

We do not by any means intend to argue that current research in social psychology is thus healthy and has sufficient self-cleansing capabilities. In fact, we agree with the previous literature that it is important to be aware of the susceptibility of current research practices to the inflation of false-positive rates (Bakker, van Dijk, & Wicherts, 2012; Fiedler, 2011; Simmons et al., 2011; Vul et al., 2009). However, we also feel that recent contributions to the debate tend to place a disproportionate emphasis on the inflation of Type-1 error rates and do not pay a similar amount of attention to the fact that other factors may reduce false-positive rates. This unbalanced discussion is unfortunate because mindless over-concern with the inflation of false-positive rates without considering such opposing factors may unnecessarily constrain research endeavours, causing researchers to underestimate the validity of their discoveries and journal editors to make unreasonable decisions. In addition, given that false-positive findings in psychology are being paid increasing attention even among the general public (e.g., Carey, 2011; Young, 2012), people outside the field of psychology may begin to consider psychology as less scientific than it really is. Our main message is that the issue of false-positive rates is dependent on the research context and that researchers should consider both factors that promote and factors that reduce false-positive rates simultaneously when evaluating the scientific validity of studies. We agree with the current criticisms that researchers should critically examine current research practices in social psychology, but we will shed light on this issue from a different angle, hoping that readers can more deeply understand the complicated nature of false-positives in research practice.

In the following, we discuss three factors common in current research practice in social psychology that may serve to preventing false-positive findings: theoretical predictions, papers with multiple studies, and additional data collection based on observed *p*-values. All three issues have to some degree been addressed in the extant methodological literature. The importance of theoretical predictions and papers with multiple studies have been discussed in the general

methodological literature (Abelson & Prentice, 1997; Rosenthal & Rosnow, 1985) and in the context of statistical power (Fiedler, Kutzner, & Krueger, 2012; Maxwell, 2004). Strategies for additional data collection have been discussed particularly in studies on sequential clinical trials (Cook, 2002; Lai, Lavori, & Shih, 2012). Nevertheless, we think it is worthwhile to examine the joint operation of these issues, as they are largely overlooked in the recent discussions on false-positives.

Theoretical Predictions

Ideally, psychological research is guided by theories, and good theories have the capability to integrate and explain a host of findings. Without theories, research findings lack integration and are difficult to interpret. Importantly, theories lead to testable predictions about the directions of effects or relationships. For example, in research on stereotype threat, it is theorized that people experience anxiety or concern in a situation where they have the potential to confirm a negative stereotype about their social group (Steele & Aronson, 1995). Based on this theory, researchers can make the prediction that when a stereotype about expected performance is made salient, individuals who belong to negatively-stereotyped groups will perform more poorly than they otherwise would (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). The plausibility of the theoretical prediction becomes stronger as empirical findings to confirm the theory accumulates. Crucially, such directional predictions force researchers to perform conservative tests, thus reducing Type-1 error rates, sometimes to a substantial degree. Imagine a researcher who wants to show a stereotype threat effect as described above using a simple 2 (stereotype threat: high vs. low) × 2 (group: negatively stereotyped vs. neutral) between-subjects factorial design. To test the theoretical prediction, this researcher would be required to show three supportive results from statistical tests: a significant interaction between stereotype threat and group, a significant simple main effect of stereotype threat for the negatively stereotyped group (in the direction of stereotype threat reducing performance), and a non-significant simple main effect of stereotype threat for the neutral group. When the true effect does not exist, the probability of obtaining a false significant interaction by chance is equivalent to the pre-determined alpha value (throughout this article we assume the alpha value is set to .05). However, there is no guarantee that the direction of this chance interaction is consistent with the theoretical prediction. Furthermore, even if the direction of the interaction happens to be consistent with the theoretical prediction, researchers need to take a step further to show statistically significant effects (or non-significant effects, when one expects a null effect within a condition) in a series of post-hoc tests. As a result, the overall false-positive rate to obtain the results that are consistent with the theoretical prediction is much lower than the nominal alpha value; it is unlikely that a significant interaction obtained by chance passes the additional post-hoc tests in the predicted directions[1].

Directional Hypothesis and Type-1 Error Rates: Statistical Simulation

To quantify this argument, we conducted a simulation study investigating Type-1 error rates under a variety of research designs and theoretical predictions (Figure 1). Specifically, for each design and theoretical prediction, we repeatedly (replications = 50,000) generated data (N = 20 for each cell) from the population without any group differences, and computed the proportion of results that showed significant effects which are consistent with the aforementioned theoretical prediction (including post-hoc tests)[2]. Because these data were generated from the null population, the proportion of these consistent results is considered to be the Type-1 error rate. For simplicity, post-hoc tests were performed with independent sample t-tests for all the cases[3]. We proceeded to post-hoc tests only when the focal omnibus test from the analysis of variance (ANOVA) was statistically significant ($p < .05$). It is possible to analytically derive Type-1 error rates for some cases. For example, the false-positive rate for a simple two-group experiment is obviously .05 when no directional hypothesis is tested (Case 1 in Figure 1). However, we conducted simulations

for all cases for the sake of comparison.

The simulation indicated that performing statistical tests according to theoretical predictions substantially reduces false-positive rates (see Figure 1). Case 2 in Figure 1, the simplest scenario, assumes an experiment of two between-subject groups with the hypothesis that the second group would score higher than the first group (but not vice versa). In this case, even if a *t*-test showed a significant difference between the groups, only half of the significant results are consistent with the prediction. Accordingly, the probability of a false-positive result that is consistent with the prediction is about 2.5%. Case 3 posits that we conducted an experiment of three between-subjects groups to test the hypothesis that the third group would score higher than the other two groups. Case 3 is typical when researchers want to investigate the effect of an experimental manipulation using two different control conditions. In this case, the overall false-positive rate of findings that were consistent with the hypothesis was less than 1%, despite the nominal alpha level of 5%.

Researchers are sometimes interested in an interaction in which an experimental manipulation is effective only for participants in one of two groups (Case 4). Stereotype threat studies are a good example. In this case, the overall false-positive rate to find the predicted pattern was also less than 1%. In some situations, researchers aim to detect a cross-over interaction in which the effect of an experimental manipulation is predicted to take opposite directions in different groups (Case 5). One example of such a cross-over interaction is research that tests a matching hypothesis, in which optimal outcomes are expected when there is congruence between personal characteristics and characteristics of the social environment (see Hunt, 1975; Higgins, 2005). In this scenario, the overall false-positive rate to obtain the predicted pattern was extremely small (less than 0.1%). These results highlight that theoretical predictions help researchers avoid the risk of committing to false-positive findings. In principle, as a priori predictions become more complex, Type-1 error rates decrease, because results have to produce a specific directional pattern and pass many post-hoc tests to be claimed as supportive evidence for the prediction.

It should be noted that directional hypothesis testing is not the only way that theoretical predictions reduce false-positive rates—theory can reduce false-positive rates in a variety of ways. Consider a common practice in which researchers conduct some additional analyses (with the same dataset) to further validate their theoretical hypothesis. For example, Legault and Inzlicht (in press) tested the hypothesis that autonomous motivation enhances neuroaffective responsiveness to failure measured by the error-related negativity (ERN). The authors experimentally manipulated autonomous motivation (Experiment 2) and found that the ERN was amplified when participants' autonomous motivation was induced by task instructions. Importantly, the authors further confirmed their hypothesis by showing that individual differences in autonomous motivation assessed by self-reported questions were positively related to those in the ERN within each group (i.e., after experimental groups were controlled). Such follow-up analyses that expect a significant effect based on a theoretical prediction would also contribute to the reduction of Type-1 error rates (especially when the follow-up analyses are independent of the original analysis), because obtaining supportive evidence in follow-up analyses is unlikely if the first analysis revealed a significant effect only by chance.

Issue of Post-hoc Hypothesis Generation

Our simulation rests on the assumption that researchers make a prediction prior to the data-analysis. It has been pointed out, however, that researchers sometimes present post-hoc hypotheses in a research report as if they were, in fact, a priori hypotheses (John et al., 2012; Okada & Shimokido, 2001; see also Fanelli, 2012). Kerr (1998) called this research practice HARKing (Hypothesizing After the Results are Known). HARKing provides researchers with possibilities to interpret unexpected findings, and potentially goes against any decrease of Type-1 error rates as

presented in Figure 1. We agree with this argument. However, although it is often possible to make post-hoc predictions for some specific patterns of unexpected results, we think it would usually be quite difficult to find satisfactory theoretical explanations for all possible patterns of results. Whereas most theories in psychology are not sufficiently precise to derive unequivocal predictions (Lilienfeld, 2012), they are not sufficiently weak to make it possible to derive *any* pattern of predictions. For example, if researchers conducting an experiment on stereotype threat observed that a stereotype threat manipulation elicited a performance decrease in the control group but not in the negatively stereotyped group, any strong claim that the researchers expected such a finding from the start would sound dubious in most cases. We are not suggesting that it is impossible to advance post-hoc hypotheses. Rather, we are arguing against the extreme view that researchers can produce strong post-hoc explanations for *any* pattern of findings[4]. Importantly, as long as certain patterns of results cannot be explained by theory, false-positive rates may still remain lower than the nominal alpha level. Therefore, even if we take into account the possibility of post-hoc hypothesizing, psychological theories can play an important role in reducing false-positive rates to some extent. The crucial point is that strong theories must be explicit about findings that cannot occur (Platt, 1964).

These considerations offer an important practical implication. To evaluate the likelihood of false-positive findings when a priori hypotheses are reported in a paper, readers (including reviewers and editors) should be sensitive to whether these hypotheses are non-arbitrary and theoretically meaningful rather than having been constructed post hoc (i.e., theoretical predictions are well validated in a paper). As previous reports indicated, researchers are generally aware that post-hoc predictions are widespread in published research (John et al., 2012; Okada & Shimokido, 2001). In spite of this awareness, it seems likely that researchers pay relatively little attention to the meaningfulness of hypotheses, as compared with the attention devoted to empirical findings. Ensuring that a theoretical prediction is made a priori, however, actually constitutes an important factor to assess whether the prediction can serve to prevent false-positives[5]. When it seems unlikely that the prediction has been made a priori (e.g., because it is inconsistent with cumulative prior evidence), the susceptibility of the respective research to false positives should be assessed through other means such as replication studies (see also the concept of confirmatory studies; Wagenmakers et al., 2012), as discussed in the next section.

Performing Multiple Studies

Diener (1998), in his editorial note in the *Journal of Personality and Social Psychology*, strongly recommended that articles should include more than a single study (see also Carver, 2004). In line with this suggestion, over the last two decades it has become increasingly common to include multiple studies in a single paper in social psychology (Sherman, Buddie, Dragan, End, & Finney, 1999; the recent trend to report single-study papers will be discussed later). There are many reasons to include multiple studies in a single paper. For example, additional studies may help researchers address potential problems identified in an initial study and thus provide stronger empirical support for their theoretical claims. Including multiple studies with different methodologies may also expand the generalizability of primary findings.

Whatever the reason, one important consequence is that multi-study papers can potentially reduce false-positive rates of findings to a substantial degree. In a multi-study paper, researchers commonly conceptually replicate their primary findings in subsequent studies (Schmidt, 2009). When the predicted effect is not true (i.e., the null hypothesis is correct), the chance of obtaining a significant effect in the first study is supposed to be 5%. Importantly, the chance of obtaining the same significant effects across two independent consecutive studies in a row in a single paper is the power of this value, i.e., $0.05 \times 0.05 = 0.0025$[6]. This number is considerably smaller than the Type-1 error rate designated for each of the two studies separately. One could argue that the false-

positive rate for each study may have been underestimated due to factors that increased Type-1 error rate (John et al., 2012; Simmons et al., 2011). But even if we assume that the Type-1 error rate is inflated up to, for example, 15%, the overall false-positive rate across the two studies is 0.15 x 0.15 = 0.025, which is smaller than the conventional alpha-level. If researchers were to consistently report more than two studies (which is already a common practice for research published in high-ranking journals in psychology), false-positive rates would further decrease exponentially. The common practice of replicating results within a single paper serves to reduce the rate of published false-positive findings.

Conceptual Replication in Multi-Study Papers

The value of replication itself has been widely acknowledged and has recently attracted increased attention from researchers (Nosek, Spies, & Motyl, 2012; Roediger, 2012; Schmidt, 2009). This is a very important step toward gaining firm ground in scientific psychology—the importance of replication to prevent false-positive findings cannot be over-emphasized. Nevertheless, this strength of multi-study research has not been sufficiently appreciated in the recent debate on false positives. One possible reason may relate to the key distinction between direct replication and conceptual replication (Nosek et al., 2012; Roediger, 2012). Direct replication attempts to reproduce a result using the same conditions, materials, and procedures as in the original study. In contrast, conceptual replication, which is more common in multi-study research, involves deliberately or systematically changing the operationalization of the key elements of the original study (such as the independent variable, dependent variable, study materials, and so on). With this distinction in mind, Simmons et al. (2011) argued that conceptual replication (within a study), unlike direct replication, does not help reducing false-positive findings because such replication does not bind researchers to make the same analytic decisions, leaving researchers leeway for capitalization on chance. Accordingly, a multi-study paper, which is typically comprised of multiple conceptual replications, is not considered as directly addressing the false-positive problem.

We agree that conceptual replication provides some room for researchers to take advantage of multiple analytic options to find a significant effect by chance. However, this argument underestimates the exponential power of replication to reduce Type-1 error rates. As suggested earlier, showing multiple replications of an effect could easily offset a small to moderate inflation in the Type-1 error rate. In addition, although researchers could have multiple analytic options for conceptual replication, these researchers' degrees of freedom are typically constrained. For example, if a researcher decided to use log-transformation in the first study, the researcher is not free to choose other transformation options in the following studies because readers (including editors and reviewers) expect some analytic consistency across the studies. As such, conceptual replication often helps reduce, if not perfectly eliminate, researchers' degrees of freedom (see also Garner, Hake, & Eriksen, 1956). We acknowledge that direct replication is a more powerful tool to prevent false-positives than conceptual replication. However, we are worried that the strong emphasis on direct replication in the recent literature may lead researchers to underestimate the value of conceptual replication in a multi-study paper. The value of conceptual replication should be appropriately judged in light of the specific context of each paper.

Two qualifications should be made. First, conceptual replications are beneficial as long as each replication has sufficient methodological quality to address the research question in focus. It is sometimes the case that papers open by presenting one or more studies having critical design problems that are later addressed by more adequate studies, as researchers think such a sequence may be more impressive overall (Bones & Johnson, 2007). In such a case, the validity of the findings strongly relies on the subset of the adequate studies, and the earlier "straw-man" studies should not be counted as replications (see Kane, 1992). Second, our argument focuses on the value

of conceptual replication *within* a multi-study paper. As Pashler and Harris (2012) point out, it may be more problematic if researchers rely only on conceptual replication to replicate the findings from studies that were conducted by other researchers. In such situations, the failure of conceptual replication may not threaten the original published findings because conceptual replication performed in another laboratory may allow for ad-hoc interpretations of failure without questioning the veracity of the original report. As a result, published false-positive findings would be likely to persist. In the context of a multi-study research project, however, failure to replicate (either with a conceptual or a direct replication) would make it difficult to publish the findings as a multi-study paper. Accordingly, in this case, conceptual replication does have some power to prevent false-positive findings. In this respect, our arguments never intend to discourage the recent large-scale direct replication attempts from independent researchers (Open Science Collaboration, 2012).

Multi-study Papers and Publication Bias

On a related note, Francis (2012a, 2012b, 2012c, 2012d; see also Schimmack, 2012) recently published a series of analyses that indicated the prevalence of publication bias (i.e., file-drawer problem) in multi-study papers in the psychological literature. In each of these analyses, Francis followed the same basic methodology: he selected a specific multi-study paper that consistently showed significant effects across all the studies reported (typically 5-10 studies), computed the statistical power (i.e., the probability of rejecting the null hypothesis when the null hypothesis is false) for each study in that paper, and multiplied the power probabilities to estimate the probability that the observed findings would all reject the null hypothesis in a row (Ioannidis & Trikalinos, 2007). The results revealed that this probability was remarkably small (e.g., 0.02) for the papers analysed, suggesting that the studies that were reported may have been conducted but that additional unpublished experiments exist that failed to obtain statistically significant effects (publication bias). Therefore, Francis argued that there may be many cases in which the findings reported in multi-study papers are too good to be true.

It is beyond the scope of this article to discuss whether publication bias actually exists in these articles or how prevalent it is in general. However, one interesting observation is that, based on this logic, multi-study papers reporting only significant effects would be more likely to indicate publication bias, because the probability that all of the observed findings would reject the null hypothesis tends to become smaller as the number of studies increases. This is especially true given that many studies in psychology are underpowered (Cohen, 1988; Sedlmeier & Gigerenzer, 1989). This observation may seem inconsistent with our proposition that findings should be considered as more reliable when based on more (successful) studies. Should multi-study findings be regarded as reliable or shaky evidence?

This seeming paradox can be resolved by distinguishing between overestimation of effect sizes and false-positive findings. A publication bias, if it exists, leads to overestimation of effect sizes because some null findings are not reported (i.e., only studies with relatively large effect sizes that produce significant results are reported). The overestimation of effect sizes is problematic, particularly when the effect magnitude is of practical concern (e.g., the effect of a medical drug). However, the presence of publication bias does not necessarily mean that the effect is absent (i.e., that the findings are falsely positive). In fact, as the number of successful studies increases, the possibility that the effect is a false-positive one becomes unlikely even when a publication bias is present. Imagine a paper that included eight experiments that showed consistent significant results. When the studies are underpowered, people may presume that the researcher may have run, but not reported, some experiments that failed to obtain statistically significant effects. However, even if this is the case (i.e., if the researcher conducted more than eight experiments), it is unrealistic that as many as eight statistically significant results were produced

by a non-existent effect. The researcher may have conducted ten, or even twenty, experiments until he/she obtained eight successful experiments, but far more studies would have been needed had the effect not existed at all.

To illustrate this point, we computed the probability of obtaining 1, 4, or 8 significant results (alpha level = .05) in *k* or less than *k* experiments when the null hypothesis is correct (Figure 2). This probability can be analytically derived by computing the cumulative probabilities of the negative binomial distribution. Figure 2 indicates that researchers would need to conduct a considerable number of experiments to report multiple significant results if the null hypothesis is correct. For example, if the effect does not exist, 100 experiments are still not sufficient to obtain eight (falsely) positive results (the probability of obtaining eight significant results in 100 experiments is less than 15%). This consideration suggests that, although experiments within multi-study papers may be susceptible to publication bias, they have the strength that a series of replications is more resistant to obtaining false-positive findings. Therefore, while it may be true that many multi-study papers in psychology are, in fact, contaminated by publication bias, publication bias should not be confused with false-positive findings. Publication bias simply means that the effect size is overestimated—it does not necessarily imply that the effect is not real (i.e., falsely positive)[7].

The most straightforward solution to this paradox is to conduct studies that have sufficient statistical power. Our argument implies that publication bias and false estimates of effect sizes should not be interpreted as indicating false-positive findings, but we do not intend to defend underpowered studies themselves.

Sequential Data Collection Contingent on Observed *p*-Values

With regard to practices that contribute to the inflation of false-positive rates, it is worthwhile to reflect upon factors related to those practices that can act counter to such an inflation. Adding data to an existing study that initially produced non-significant results is an example for a research practice that increases false-positive rates, but some realistic factors can prevent an increase that is excessively large.

Imagine that you conducted an experiment with two groups (N = 20 per cell) and obtained a p-value of .07 when conducting a t-test. Given the conventional standard in psychological journals, the data do not seem publishable. You might then be tempted to collect 10 more participants per cell, because the sample size may not have been large enough to detect a significant effect. However, recent papers have questioned such practice, because adding participants this way could increase false-positive rates (John et al., 2012; Simmons et al., 2011). Indeed, research in methodology has long indicated that sequential data collection procedures increase false-positive rates (Jennison & Turnbull, 1990; McCarroll, Crays, & Dunlap, 1992; Strube, 2006). A simulation by Simmons et al. (2011) showed that such sequential data collection may result in a substantial inflation of Type-1 error rates, sometimes as high as more than 20%.

As reported by John et al. (2012), additional data collection (when the results are not significant) seems to be one of the most common research practices in psychology. Given the pervasiveness of this practice, Type 1 error rates of 20% as illustrated by Simmons et al. (2011) may be shocking to many researchers. Researchers who employed a sequential data collection procedure may be worried that their findings are likely to be false-positives. Researchers may also see the field as a whole full of false-positive findings. Researchers may think that adding data when results did not reach significance is a completely flawed practice and should be banned for any reasons. Before jumping to such simple and extreme conclusions, however, we need to scrutinize how, and to what extent, such practices really increase Type-1 error rates, to have a more judicious/balanced view on this issue

Informational Value of Additional Data

When an experiment produces a non-significant effect, this does not imply that the effect is null. Rather, non-significant results merely indicate that the observed data are not fully inconsistent with the null hypothesis. Indeed, when an effect is close to significant (e.g., $p = .07$), the observed data are likely to be more consistent with an alternative hypothesis than with the null hypothesis. Take the example considered above. We can compare how consistent the observed data are with the null versus an alternative hypothesis by computing a likelihood ratio (using the noncentral *t*-distribution). With a medium-sized effect (Cohen's $d = .50$) as the alternative hypothesis, the obtained likelihood ratio is 5.17, indicating that the alternative hypothesis is five times more likely than the null hypothesis[8] (Pawitan, 2001). Although this is not conclusive evidence that the alternative hypothesis is correct, it is reasonably high enough to motivate researchers to collect more data (see Jeffreys, 1961).

Importantly, a decision to add more data has merit despite an increase in Type-1 error rates: Increased precision (e.g., narrower confidence intervals). By adding data, we can estimate true effect sizes more reliably. In fact, if the null hypothesis is correct, adding data is more likely to move *p*-values away from the threshold (i.e., .05), whereas the opposite is true if an alternative hypothesis is correct. To demonstrate this point, we conducted a simulation (replications = 200,000) to estimate the conditional probability of obtaining a significant effect by adding data (once) when the initial *p*-value was close to significant (i.e., $.05 < p < .10$). The simulation was performed by changing a variety of parameters, including initial sample size ($N_{initial}$), number of added participants (n_{added}), and population effect size (in Cohen's *d*). The findings (see Table 1) revealed that, when the *p*-value was close to significant, adding data was much less likely to produce a statistically significant result if the null hypothesis was correct (13-23% with $d = 0.00$) than if an alternative hypothesis was correct (e.g., 49-77% with $d = 0.50$). Note that we used the number of replications that had an initial *p*-value between .05 and .10 as the denominator to compute the probabilities; Therefore, this probability of 13-23% does not represent Type-1 error rate.[9] Clearly, additional data contain information that helps researchers distinguish whether a true effect exists or not.

Type-1 Error Rate and Adding Data Only When Results Are Promising

Of course, adding data inevitably increases Type-1 error rates. In the case of adding participants only when results are marginal ($.05 < p < .10$), however, the combination of (1) the greater plausibility of the alternative hypothesis and (2) the higher precision of estimates after adding data, leads to a relatively small increase in false positives, which may be less problematic. To illustrate this point, Figure 3 plots the results of another simulation (replications = 20,000) that used a procedure identical to the sample simulation conducted by Simmons et al. (2011, Figure 1). The simulation posits an experiment with two between-participants groups and that a *t*-test has been conducted to test whether there is a significant difference between the two groups. The data were generated from the population with no between-group differences, and, therefore, claimed significant effects were all counted as false-positives. Dashed lines represent the direct replication of Simmons et al. (2011), depicting the false-positive rates when a researcher who had already collected either 10 (circled points) or 20 (squared points) observations within each of two conditions again tested for significance every 1, 5, 10, or 20 per-condition observations after the initial test. Data collection was stopped either once statistical significance was obtained or when the number of observations in each condition reached 50. The results replicate the findings of Simmons et al. (2011), showing that such sequential data collection substantially inflates Type-1 error rates, sometimes as high as more than 20%.

However, our simulation also revealed that this is not always the case. Solid lines represent the false-positive rates if data are added only when the current result is promising (i.e., $.05 < p < .10$). Under this practically realistic condition (and this is the case that we highlighted in Table

1), the Type-1 error rate was considerably smaller (below 7.1%) than when adding data irrespective of current results. This finding indicates that collecting additional data guided by the observed *p*-value may prevent the inflation of Type-1 error rate, sometimes to a considerable degree. Observed *p*-values have some informational value, and assuming that our decision to collect more data is influenced by the observed *p*-values (and we believe this assumption pertains to a more realistic scenario in practice), the effects of sequential data collection on Type-1 error rate inflation may not be as consequential as reported by Simmons et al. (2011).

Although the simulation showed that the inflation of Type-1 error rates is substantially reduced in a realistic situation, it is true that the empirical Type-1 error rate is still above the nominal alpha value (.05). Some researchers may argue that even such a slight inflation should not be allowed and that we should stick to the conventional standard. Another, more elaborated reaction might be that it would be necessary to adjust *p*-values to achieve the nominal Type-1 error rate (see Cook, 2002; Lai, Lavori, & Shih, 2012), even if the inflation is small. We have three responses to this question. First, our main message is that the susceptibility to false-positives should be evaluated based on the specific context of the research, jointly considering all possible factors enhancing and preventing the likelihood of false positives. Therefore, the judgment on whether this deviation is permissible or not should be made after deliberately considering all of these factors. Second, in reality, Type-1 error rates are influenced by myriad of extraneous factors that are inherent in data, as statistical tests are based on a certain set of assumptions (e.g., normality, equal variance, sample independence). In practice, we rarely, or sometimes simply are unable, to test these assumptions. As a result, empirical Type-1 error rates always (at least slightly) deviate from the nominal Type-1 error rate. Accordingly, Bradley (1978), for example, argued that a test could be considered robust to the violation of assumptions if its empirical rate of Type 1 error is below .075. In the particular cases we simulated, the empirical Type-1 error rate was within this limit. Third, our motivation to perform the simulation was not to argue that collecting additional data should generally be recommended, but simply to show that the consequences of this practice have been overstated.

General Discussion

We discussed three factors common in current research practices in social psychology that may contribute to preventing false-positive findings. First, we showed that theoretical predictions may decrease Type-1 error rates, often to a considerable degree. Second, we pointed out that multiple-study papers involving a replication of findings, which is a current standard in social psychology, can also contribute to the reduction of false-positive findings. Third, we suggested that sequential data collection may not be as deteriorating as illustrated by previous work, provided a realistic scenario where researchers make use of the information in the observed *p*-values rather than thoughtlessly gathering additional data.

Practical Implications

Being aware of factors preventing false-positives, in addition to the facilitative factors discussed in previous research, has several implications. First, people can have a more balanced view on the pervasiveness of false-positives in social psychology. Recent concern about questionable research practices (Simmons et al, 2011; John et al., 2012) has produced a strong scepticism towards the credibility of past findings in social psychology, and even psychology in general. Given that the recent debate only focused on factors that facilitate false-positives, however, it is possible that many people (both within and outside the field) have exaggerated or extreme impressions about the pervasiveness of false-positives in social psychology (e.g., "researchers can present anything as statistically significant"). When considering both facilitative and preventive factors, people can evade such extreme views. Our simulation on the sequential data collection procedure that incorporated a realistic decision factor, for example, clearly makes

this case.

Second, the preventive factors discussed in this paper may help readers (including editors and reviewers) to gauge the susceptibility of reported research to false positives in fair and pragmatic ways. Imagine a paper that reports two experiments showing a consistent pattern of interaction. Given that most of the factors that increase false-positives are typically invisible to readers, it is easy to be sceptical about the reliability of the results. Accordingly, an easy solution for editors and reviewers would be to demand an additional experiment replicating the interaction. However, our research suggests that the false-positive rate for this paper would be already sufficiently small and preventive factors would be likely to outweigh any negative effects of questionable research practices, as the researcher(s) already conducted two studies that confirmed the same interaction. Thus, in this particular example, requiring additional replication based on the concern about false-positives may be unreasonable and demand unnecessary additional research effort (see also the issue of statistical power discussed below).

As another example, if a reviewer or an editor reads a paper reporting a marginally significant effect for a primary study hypothesis (e.g., $p = 0.09$), recent concerns about sequential data-collection procedures may motivate the reviewer or editor not to require additional data collection but to conduct an additional study to avoid false-positives. However, considering preventive factors, the reviewer or editor may make a more contextualized decision. One reasonable action would be to first consider other factors that may have decreased the likelihood of false-positives (e.g., whether the authors have tested a hypothesis that has a relatively complicated directional pattern), and, if they are present, ask the author(s) to collect more data, because the (presumably small) negative effect of this particular additional data collection should be overridden by the other preventive factors. If the effect became significant after adding data in a planned way, researchers and editors could be reasonably confident about the validity of the focal hypothesis. As these examples indicate, adopting a balanced view on the false-positive issue could help researchers evade the burden of conducting many unnecessary experiments.

Related Issues

The current paper focused on three factors that prevent the inflation of false-positive rates. There are, however, other essential issues that are worth mentioning. Among these issues, we consider three: Sample size, single-study papers, and false-negatives.

It is not common to collect a sufficiently large number of participants for behavioral experiments in social psychology. Including larger samples is indeed another way to prevent an inflation of false-positive rates. At the same time, large sample sizes also decrease false-negative rates (addressed below) and increase the precision of effect size estimates. Accordingly, there is no doubt that studies involving large samples should generally be preferred over small-sample studies. As long as involving more participants does not jeopardize the representativeness of the sample, it is always advisable to collect as many participants as possible. However, one important qualification that we should correctly understand is that small-sample research does not by itself increase false-positive rate. What small sample research increases is the susceptibility to factors that increase false-positive rate (e.g., sequential data collection; see also Button et al., in press, for other important issues associated with small sample studies). In other words, when it is clear that researchers did not engage in questionable research practices (see Simmons, Nelson, & Simmonsohn, 2012) and still obtained a significant effect, the false-positive rate is always below the nominal rate (e.g., 0.05) regardless of sample size, and small sample size would not be a cause of possible false-positives[10]. Therefore, the reliability of a finding may not be questioned only because the sample size is small. For small-sample studies as well, a contextualized assessment is needed to evaluate the susceptibility to false-positives.

Contextualized evaluation should also be considered for single-study papers. Recently,

contrary to the conventional preference for multi-study papers, the number of journals that publish single-study papers (sometimes in the form of *Brief Reports*) has rapidly increased. Whereas single-study papers have appeal in that they allow for fast data collection and quick publications, our analysis indicates that these papers are more susceptible to the inflation of false-positives due to lack of replication (see Ledgerwood & Sherman, 2012, for a more comprehensive discussion). In addition, if published as brief reports, space limitations may provide a rationale not to report questionable research practices that increase false-positive rates. Therefore, as with small-sample studies, single studies by themselves do not increase false-positive rates, but increase the susceptibility to factors that boost false-positives. We do not have a strong opinion on the pros and cons of single-study papers, but in light of the current paper, one point that has not been made in the ongoing discussion (Bertamini & Munafo, 2012; Ledgerwood & Sherman, 2012) is that researchers should pay more attention to other factors than the number of studies that prevent false-positives to make an overall assessment of a paper's credibility. For example, if a paper with a single experiment tested a hypothesis with two independent analyses (as described earlier), we could substantially discount the possibility of false-positives in confirming the hypothesis.

The current paper focused on false-positives, but false-negative findings (Type II errors) warrant attention as well. False-negatives represent failures to reject the null hypothesis when the alternative hypothesis is correct. Researchers have argued that such failures may reduce the chances for scientific discovery, and that neglect of false-negatives can limit novel and generative research endeavours (see Fiedler et al., 2012; Lieberman & Cunningham, 2009). Importantly, false-positive and false-negative rates are negatively interdependent when sample size and effect size are held constant. Therefore, research practices that reduce false-positive rates, as discussed in the current paper, could potentially contribute to an increase of false-negative findings (except for increasing sample size). Accordingly, researchers should be aware of the costs and benefits of these practices. We also suspect that such potential inflation in false-negative rates inherent in psychological research may partly be linked to the motivation for researchers to employ various research tactics to achieve a significant *p*-value (i.e., $p < .05$) outlined by Simmons et al. (2011). In this respect, we agree with Simmons et al. (2011) that editors and reviewers should be more tolerant of imperfections in results. In cases such as those we illustrated (e.g., studies expecting a specific pattern of interaction or studies with many replications), it is very difficult to obtain perfect results that are all significant, unless the effect size or the sample size is unusually large (see Maxwell, 2004). We are concerned that undesirable research practices will persist unless editors and reviewers become aware of the fact that in some cases (again, we emphasize dependence on the research context) conventional journal requirements impose unreasonable expectations on researchers.

Concluding Remarks

To reiterate, in discussing factors that can prevent an inflation of false-positive rates, we do not argue that psychologists should not be worried about false-positive findings. We fully agree that researcher degrees of freedom (and other incentive factors, as discussed by Nosek et al., 2012) can inflate Type-1 error rates and that some practices may have led to publications that actually report false-positive findings. In fact, social psychology and the behavioral and social sciences more generally have witnessed many findings that became famous but are difficult to replicate. However, it is also important to consider false-positive issues in the context of individual studies by taking the specific features of individual study designs into account and by considering both factors that enhance and factors that reduce the risk of obtaining and reporting false-positive findings. Such a balanced view could help prevent an unnecessary devaluation of psychological findings and pave the way for a more productive discussion on how to make reliable and innovative scientific discoveries in the field.

References

Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypothesis. *Psychological Methods*, *2*, 315-328.

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., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678. doi: 10.3758/s13428-011-0089-5

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

Bertamini, M., & Munafo, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science*, *7*, 67-71. doi: 10.1177/1745691611429353

Bones, A. K. & Johnson, N. R. (2007). Measuring the immeasurable, or, "Could Abraham Lincoln take the implicit association test?" *Perspectives on Psychological Science*, *2*, 406-411.

Bradley, J. V. (1978). Robstness? *British Journal of Mathematical & Statistical Psychology*, *31*, 144-152.

Button, K. S., Ionnidis, P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (in press). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*.

Carey, B. (2011, November 3). Fraud case seen as a red flag for psychology research. *The New York Times*, A3.

Carver, C. S. (2004). Editorial. Journal of Personality and Social Psychology, 86, 95.

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

Cook, T. D. (2002). P-value adjustment in sequential clinical trials. Biometrics, *58*, 1005-1011.

Diener, E. (1998). Editorial. *Journal of Personality and Social Psychology*, *74*, 5-6.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association*, *50*, 1096-1121.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891-904. doi: citeulike-article-id:9798745

Fiedler, K. (2011). Voodoo correlations are everywhere-Not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163-171. doi: 10.1177/1745691611400237

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from alpha-control to validity proper: Problems with a shortsighted false-positive debate. *Perspectives on Psychological Science, 7*, 661-669.

Francis, G. (in press). Replication, statistical consistency, and publication bias. Journal of Mathematical Psychology.

Francis, G. (2012a). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1203591109

Francis, G. (2012b). Publication bias in "Red, rank, and romance in women viewing men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*.

Francis, G. (2012c). The same old New Look: Publication bias in a study of wishful seeing. *I-Perception*, *3*, 176-178.

Francis, G. (2012d). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151-156. doi: 10.3758/s13423-012-0227-9

Gamer, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationalism and the concept of

perception. *Psychological Review*, *63*, 149-159.
Garcia-Perez, M. A. (2012). Statistical conclusion validity: some common threats and simple remedies. Frontiers in Psychology, *3*, 325. doi: 10.3389/fpsyg.2012.00325
Higgins, E. T. (2005). Value from regulatory fit. *Current Directions in Psychological Science*, 14, 209-213.
Hunt, D. E. (1975). Person-environment interaction: A challenge found wanting before it was tried. *Review of Educational Research*, *45*, 209-230. doi: 10.2307/1170054
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi: 10.1371/journal.pmed.0020124
Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245-253. doi: 10.1177/1740774507079441
Jeffreys, H. (1961). *Theory of probability (3rd Ed.)*, Oxford: Oxford University Press.
Jennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of mdedical trials: A review and commentary. *Statistical Science*, *5*, 299-317.
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.
Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*, 196-217. doi: 10.1207/s15327957pspr0203_4
Lai, T. L., Lavori, P. W., & Shih, M. C. (2012). Adaptive trial designs. In P. A. Insel, S. G. Amara & T. F. Blaschke (Eds.), *Annual Review of Pharmacology and Toxicology*, Vol 52 (Vol. 52, pp. 101-110). Palo Alto: Annual Reviews.
Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, *7*, 60-66. doi: 10.1177/1745691611427304
Lee, P. M. (1997). *Bayesian statistics*. New York: Wiley.
Legault, L., & Inzlicht, M. (in press). Self-determination, self-regulation, and the brain: Autonomy improves performance by enhancing neuroaffective responsiveness to self-regulation failure. *Journal of Personality and Social Psychology*
Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience*, *4*, 423-428. doi: 10.1093/scan/nsp052
Lilienfeld, S. O. (2012). Public skepticism of psychology: Why many people perceive the study of human behavior as unscientific. *American Psychologist*, *67*, 111-129. doi: 10.1037/a0023963
Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163. doi: 10.1037/1082-989x.9.2.147
McCarroll, D., Crays, N., & Dunlap, W. P. (1992). Sequential ANOVAs and Type I error rates. *Educational and Psychological Measurement*, *52*, 387-393. doi: 10.1177/0013164492052002014
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*.
Okada, T., & Shimokido, T. (2001). The role of hypothesis formation in psychological research. In K. Crowley, C. D. Schunn & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 445-464). Mahwah, NJ.: Lawrence Erlbaum Associates Publishers.
Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the

reproducibility of psychological science. *Perspectives on Psychological Science*, 7, 657-660. doi: 10.1177/1745691612462588

Pashler, H. & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531-536.

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

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford: Clarendon Press.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347-353.

R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Roediger, H. L. (2012). Psychology's woes and a practical cure. *APS Observer*, 25(2).

Rosenthal, R., A. & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

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. doi: 10.1037/a0015108

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309-316. doi: 10.1037//0033-2909.105.2.309

Sherman, R. C., Buddie, A. M., Dragan, K. L., End, C. M., & Finney, L. J. (1999). Twenty years of PSPB: Trends in content, design, and analysis. *Personality and Social Psychology Bulletin*, *25*, 177-187. doi: 10.1177/0146167299025002004

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

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. doi: 10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. SSRN eLibrary. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a,b,c,d,e,f). *Perspectives on Psychological Science*, *7*, 597-599

Spellman, B. A. (2012a). Introduction to the special section: Data, data, everywhere ... Especially in my file drawer. *Perspectives on Psychological Science*, *7*, 58-59. doi: 10.1177/1745691611432124

Spellman, B. A. (2012b). Introduction to the special section on research practices. *Perspectives on Psychological Science*, *7*, 655-666.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4-28. doi: 10.1006/jesp.1998.1373

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797-811. doi: 10.1037/0022-3514.69.5.797

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, *38*, 24-27.

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. doi: 10.1111/j.1745-6924.2009.01125.x

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632-638. doi: 10.1177/1745691612463078

Yong, E. (2012). Bad copy. *Nature*, *485*, 298-300.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*, e201. doi: 10.1371/journal.pmed.0050201

Footnote

1. It should be noted that there are several statistical methods that may be suited to directly test directional hypotheses, such as planned contrast analysis (Rosenthal & Rosnow, 1985) and some multiple comparison methods (e.g., Dunnett's method; Dunnett, 1955). We do not consider these methods in the current manuscript, as performing standard ANOVA-type omnibus analyses followed by post-hoc tests (Fiedler, Kutzner, & Krueger, 2012) is a conventional practice that is still considered appropriate by most behavioral and social scientists (including editors and reviewers).
2. All of the simulations reported in this article were performed using R version 2.14.2 (R Development Core Team, 2012). We made sure that the simulation results were almost the same when we reran another batch with the same number of replications. Also, in this particular simulation, the results did not depend on the *N*s for each cell.
3. We did not correct for inflation of alpha levels in multiple comparison tests in the simulation. Note that conducting multiple comparison tests would decrease Type-1 error rate, further strengthening our argument that theoretical predictions can substantially reduce false-positive rates.
4. Note that directional hypotheses are more of an issue when researchers publish multiple studies in a single paper (i.e., multi-study paper). In that case, it is extremely difficult (if not impossible) to find satisfactory explanations for effects in random directions across all the studies. The next section will discuss a topic involving multi-study papers, but for purposes of simplicity, we do not take into account the issue of theoretical predictions. However, it is important to keep in mind that theoretical predictions play a stronger role in such multi-study papers.
5. Post-hoc predictions do not decrease false-positive rates, but it is also worth noting that post-hoc predictions by themselves do not increase false-positive rates above the nominal Type-1 error rate either. It is only when researchers take advantage of factors promoting false positives that the issue of post-hoc prediction becomes especially problematic.
6. A common misconception is that the illustrated multiplication does not hold because experiments from the same paper may not be independent. However, the illustrated multiplication actually holds as long as we are concerned with the joint false-positive rate given the shared procedure and participant characteristics. In other words, two studies in the same paper can be considered as conditionally independent (i.e., these studies are independent given the shared procedure and participant characteristics). However, when we want to compute the joint false-positive rate across multi-study papers, each individual study is no longer independent because participants and experimental procedure are more similar within than across studies. The current paper focuses on the false-positive rate within a multi-study paper. Note, however, that this computation does not hold when researchers do not report non-significant experiments (see our discussion on publication bias).
7. The importance of effect size and false-positives varies depending on the research context. When investigating effects that are practical and consequential (e.g., medical trials), overestimation of effect size is especially problematic. On the other hand, when the emphasis is placed on a new discovery, extra care must be taken for false-positives.
8. When the population effect size is not a fixed value but an unknown quantity with a prior, the likelihood ratio is equivalent to the Bayes factor, which allows us to evaluate the change in the likelihood of the null and the alternative hypotheses given data. The remarkable advantage of a Bayesian perspective is that decision guided by Bayes factor is indifferent to data collection procedures (Lee, 1997; Wagenmakers, 2007).
9. When the population effect size is 0, the probability of obtaining a *p*-value between .05 and .10

is 5%. In other words, out of the 200,000 replications, there were approximately 10,000 replications that showed a marginally significant effect ($.05 < p < .10$) and 10,000 replications that showed a significant effect ($p < .05$). Of these 10,000 replications exhibiting the marginally significant effect, about 13-23% (i.e., 1,300 – 2,300 replications) showed a *p*-value below .05 with an additional data. Therefore, assuming that 13% of the marginally significant effects became significant, if one wanted to compute Type-1 error rate in this situation, Type-1 error rate would be about [10,000 (false-positive findings in the original test) + 1,300 (false-positive findings with the additional data)] / 200,000 (total replications) = 5.7%. This is the basis of computing false-positive rates in the next simulation.

10. This argument is correct when researchers use exact test statistics (such as F-tests in ANOVA or regression analysis). If researchers are interested in statistics based on an asymptotic distribution (e.g., chi-square test for a frequency table or z-test for path coefficients in structural equation modelling), sample size to some degree impacts false-positive rates. Another potential limitation of a small-sample study is that it has less statistical power to detect the violation of the assumption of a statistical test (e.g., equal variance assumption in ANOVA), although it is not very common to test statistical assumptions even in large sample studies.

Table 1

Conditional Rate of Obtaining a Significant Effect by Adding Data When the Initial p-Value is Close to Significance (.05 < p < .10).

Initial sample size ($N_{initial}$)	Added sample size (n_{added})	Population effect size (in Cohen's *d*)			
		d = 0.00	*d* = 0.30	*d* = 0.50	*d* = 0.80
$N_{initial}$ = 10	n_{added} = 5	0.218	0.348	0.488	0.674
	n_{added} = 10	0.170	0.361	0.552	0.797
	n_{added} = 20	0.125	0.404	0.663	0.918
$N_{initial}$ = 20	n_{added} = 5	0.227	0.378	0.518	0.695
	n_{added} = 10	0.204	0.424	0.609	0.830
	n_{added} = 20	0.165	0.477	0.730	0.944
$N_{initial}$ = 40	n_{added} = 5	0.227	0.386	0.518	0.702
	n_{added} = 10	0.223	0.463	0.639	0.846
	n_{added} = 20	0.194	0.533	0.766	0.952

Note. The results are displayed as a function of initial sample size ($N_{initial}$), number of added participants (n_{added}), and population effect size (in Cohen's *d*). Simulations were performed with 200,000 replications for each condition. We then selected the replications that had an initial *p*-value between .05 and .10 and investigated whether the additional data to such cases led to statistically significant results or not. As such, the number of selected replications was used as the denominator in the computation of the rate of significant effects. The numerator was the number of significant effects obtained from significance tests that were performed after the planned number of participants (n_{added}) had been added.

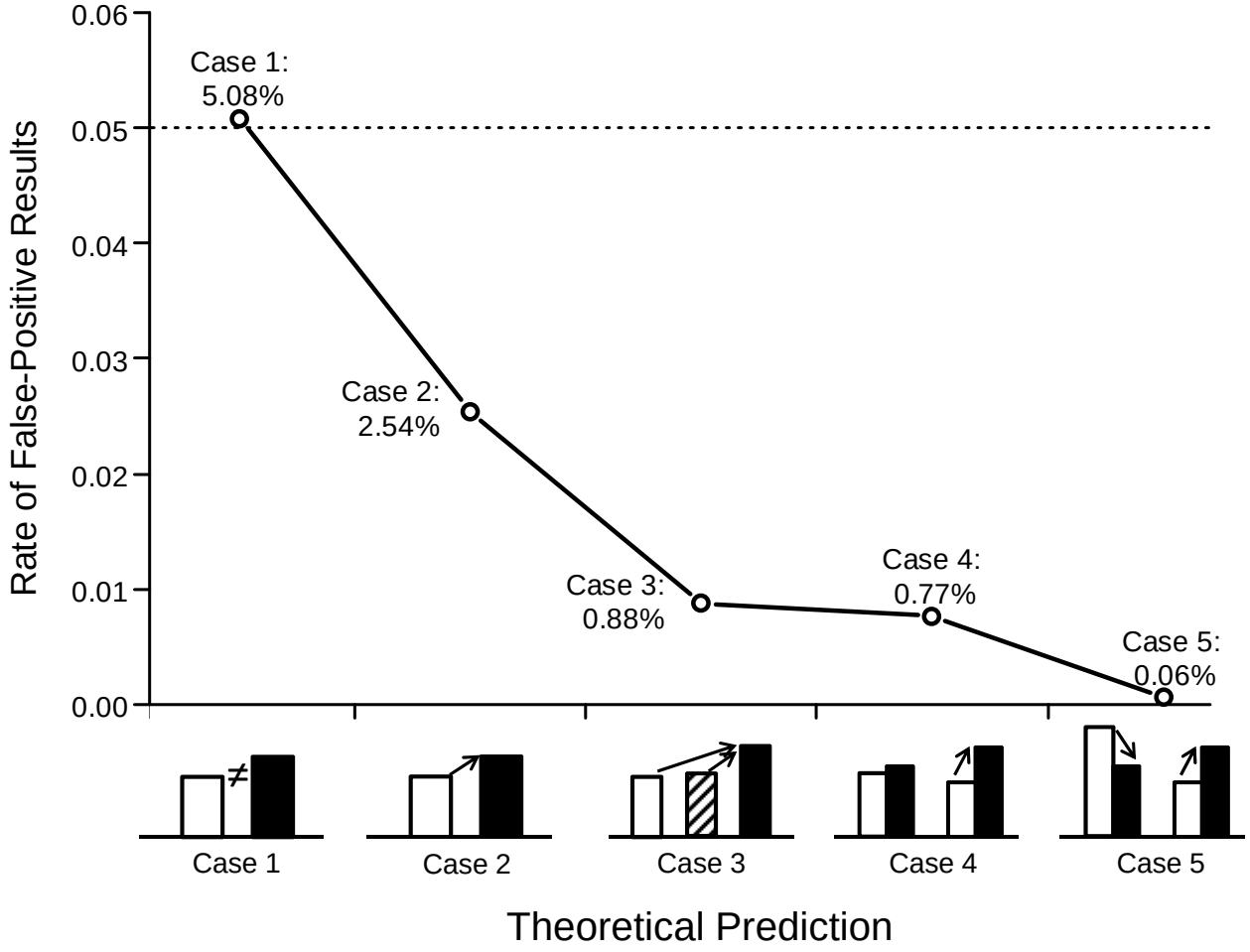

Figure 1. Probability of obtaining a false-positive result that is consistent with a theoretical prediction (based on simulation; replication = 50,000 for each case; $n = 20$ for each condition). Bar graphs below the x-axis (Case 1 – Case 5) represent the theoretical prediction, with arrows indicating the direction of the predicted results. ≠ stands for the expectation of significant difference without a directional hypothesis. For example, in Case 2, the second group (black bar) is predicted to be higher than the first group (white bar). In Cases 3 to 5, we first conducted an analysis of variance and proceeded to post-hoc tests only when the focal omnibus test (i.e., main effect or interaction effect) was statistically significant ($p < .05$). The conventional criterion ($\alpha = .05$) is highlighted by the dotted line.

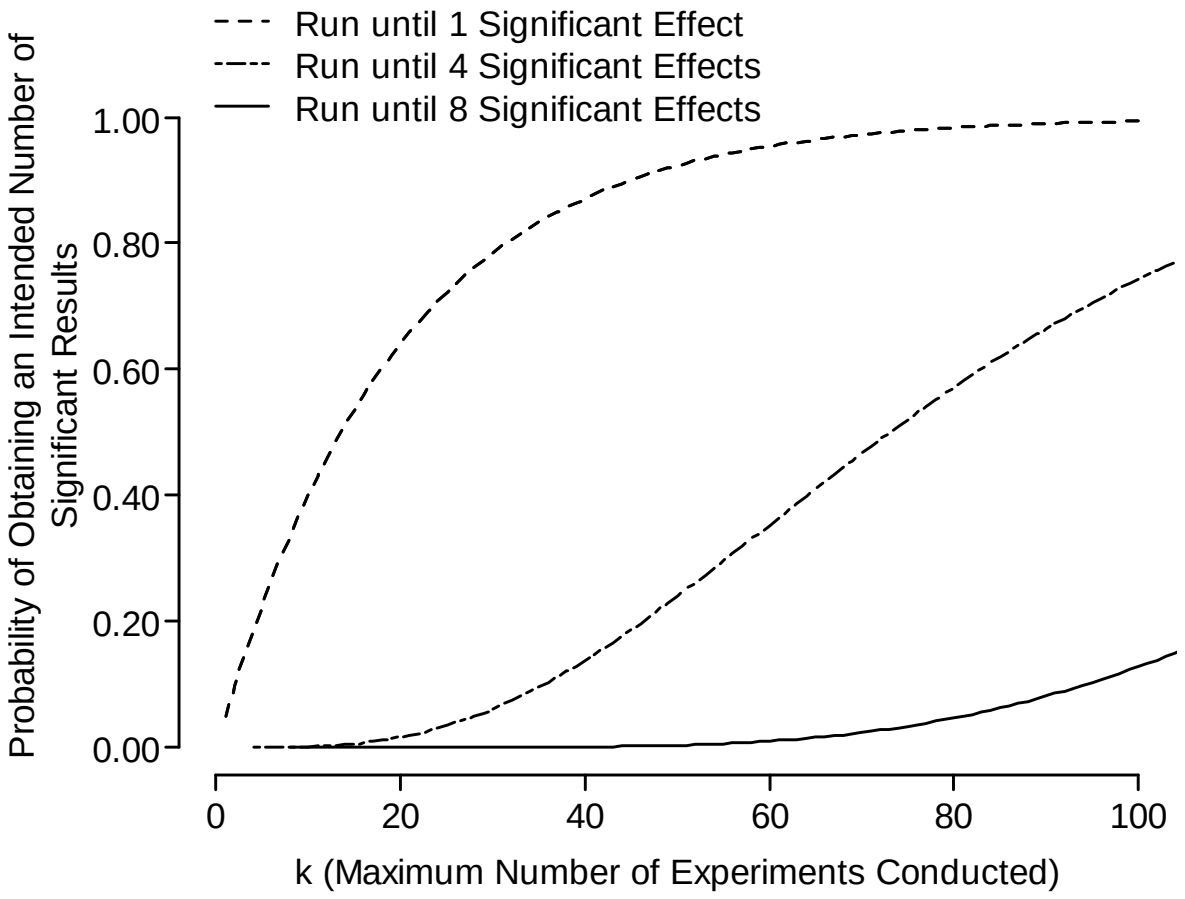

Figure 2. Probability of obtaining an intended number (dotted = 1, dashed = 4, or solid = 8) of significant results in *k* or less than *k* experiments when the null hypothesis is correct.

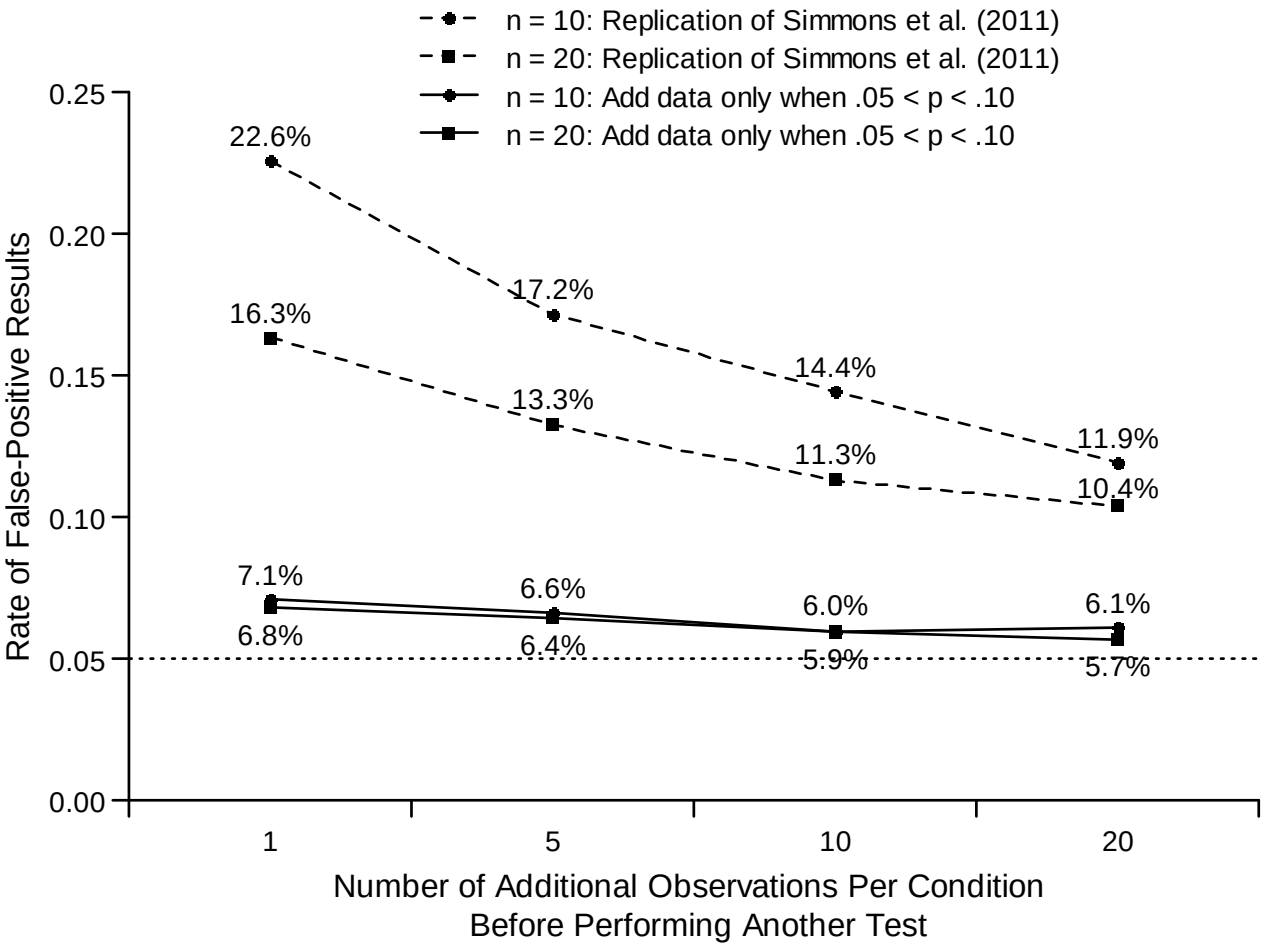

Figure 3. Probability of obtaining a false-positive result when data collection ends upon obtaining significance after sequentially collecting data and testing significance. Probability is displayed as a function of the frequency with which subsequent *t*-tests are performed (based on 20,000 replications per condition). *n* represents the two minimum sample sizes used to start performing the *t*-tests. The dashed lines show the results of the direct replication of Simmons et al.'s (2011) simulation (i.e., additional data are collected irrespective of the *p*-value of the current result). The solid lines show the results when further data were only added if the current result looked promising (i.e., $.05 < p < .10$). When the current result did not fulfil this criterion, data collection was stopped. The conventional criterion ($\alpha = .05$) is highlighted by the dotted line.