"Gee, $P = .06$; Let's Run Two More Participants!": Wicked Effects of Motivated Conditional Data Collection on False Positives in Null-Hypothesis Significance Testing

Marc Jekel
University of Cologne

It has been shown in simulation studies that the probability of a false positive research result increases rapidly when additional data is repeatedly collected and tested for statistical significance. I extend this scenario, also known as optional or conditional stopping, by a less extreme but more realistic scenario: Researchers are likely motivated to run additional participants when the initially observed p-value is close to the desired significance level. Results from two Monte-Carlo simulations show the consequences of such behavior: The probability of a false positive result increases up to 40% for studies in which researchers run additional participants and the α-error of a study can increase by 45%. Potential false beliefs that rationalize this behavior are identified and discussed.

Keywords: Optional stopping, false positive, α-error, null-hypothesis significance testing

A survey of 2,000 psychologists has shown an alarmingly high prevalence of questionable research behavior like failing to report all dependent measures in a study or selectively reporting studies that "worked" (John, Loewenstein, & Prelec, 2012). Collecting more data after determining that results are statistically non-significant has the second highest estimated prevalence of 72%. This behavior is known as optional or conditional stopping in data collection (Feller, 1940; Wagenmakers, 2007) and is highly problematic: The probability of a false positive (i.e., H_0 being true but $p < .05$) increases rapidly and leads to a statistically significant result with certainty when researchers keep running additional participants and testing for significance until a p-value below the desired significance level is acquired.

The scenario of optional stopping is unrealistic in two respects. It is unlikely that researchers *always* run more participants when the initial p-value is above the desired significance level. It is also unlikely that researchers repeatedly test for significance for each additional number of participants run—up to 30 times in the example above—without feeling guilty of performing questionable research behavior. Considering the high percentage of an estimated 72% of researchers who run additional participants after observing a statistically non-significant test result and assuming that most of those researchers do not intentionally practice behavior that damages the research field, it is more likely that researchers show less extreme but only apparently less problematic behavior that can be easily rationalized by a set of false beliefs.

Running additional participants is most likely motivated by the distance of the initial p-value to the desired significance level observed for the initial number of participants. Researchers are most likely tempted to run additional participants when observing a low p-value (e.g., .06 versus .40). A researcher may also falsely believe that running few (e.g., 2) additional participants is not problematic, at least when the initial number of participants run is high (e.g., 40 versus 20). A researcher may also falsely believe, and thereby feel justified to run additional participants, that a low initial p-value is more likely under H_1 than under H_0; s/he may also falsely believe that this is even more true for well-designed studies with (e.g.) a powerful design (e.g., within- versus between-subjects), excellent psychometric properties of the measure for the dependent variable and therefore a low error variance, and/or a high number of participants which results in a low probability for an admission error of a nonexistent effect (i.e., α-error) *and* a low probability for an omission error of an existent effect (i.e., β-error).

I label this behavior motivated conditional data collection because running additional participants is motivated by a low initial p-value. I label the effects of motivated data collection wicked because the constellation of motivational factors and false beliefs as described above maximizes the probability of a false positive. As I will show in two Monte-Carlo simulations, the probability of a false positive is highest at 40% for studies that produce an initial p-value *close* to the desired

This is a non peer-reviewed manuscript (last updated July 11, 2019). Correspondence concerning this article should be addressed to Marc Jekel, Social Cognition Center Cologne, University of Cologne, Richard-Strauss-Straße 2, 50931 Köln, Germany, E-mail: marc.jekel@uni-koeln.de.

significance level when only *few* additional participants are run. Increasing the number of initial participants further amplifies the probability of a false positive. An initial p-value close to the significance level also does not necessarily signal H_1 being true. Especially for well-designed studies with α-error = β-error = .05, an initial low p-value is even more likely to occur under the H_0 than under the alternative H_1.

In the light of the debate on the replicability of research results and false positive results due to the inappropriate application of null-hypothesis significance testing (Francis, 2012; Fuchs, Jenny, & Fiedler, 2012; Gadbury, 2012; Ionnaidis, 2005; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), it is instructional to sensitize researchers to the consequences of questionable research behavior. Additionally, it is also instructional for developing interventions to identify potential false beliefs that rationalize this behavior in the first place.

Scenario 1: Discrete motivation for conditional data collection

In the first scenario, I simulate a researcher with a discrete motivation to run additional participants. This researcher always runs a specific number of additional participants after observing a p-value below a specific threshold close to the conventional significance level of $p = .05$ for the initial number of participants and never runs additional participants otherwise. If the initial p-value drops below the desired significance level after running additional participants once, the study is counted as statistically significant. Otherwise, no additional participants are run and the study is counted as non-significant.[1]

Method

Conditions. I simulated one-sample research studies by randomly drawing for each participant one value from a normal distribution with $\mu = 0$ and $\sigma = 1$ for the dependent variable. This means that H_0 is true; there is no effect in the population. I applied for each study a one-sided t-test, expecting a treatment effect significantly ($p < .05$) greater than a specific value (e.g., 0). I varied the number of participants n_i^{ini} initially run with $n_i^{ini} \in N^{ini} = \{20, 40, 70\}$, the range of critical p-values p^{crit} for which additional participants are run with $.05 \leq p^{crit} \leq x_i \in X = \{.06, .07, \ldots, .14, .15\}$, and the number of additional participants n_i^{add} run with $n_i^{add} \in N^{add} = \{2, 5, 8, 10, 12, 15\}$. Results are therefore based on 3 (number of initial participants) ×10 (range of critical p-values) ×6 (number of additional of participants) ×500,000 (number of simulated studies) = 90,000,000 data points. Simulations were run with the software package R (2013).

Dependent variables. The first dependent variable—the probability of a conditional false positive—is the probability of the event e^{sign} that the data in a study produces a p-value below the conventional significance level of .05 although H_0 is true (1) given the event e^{crit} of observing p^{crit} below an upper bound of x_i, (2) given a number of initial participants n_i^{ini} run, (3) and given the number of additional participants n_i^{add} run, or: $p(e^{sign}|e^{crit}, x_i, n_i^{add}, n_i^{ini}, H_0)$. The second dependent variable is the increase in the α-error of a study $ratio_\alpha$ due to motivated conditional data collection, or: $ratio_\alpha = \alpha_{cond}/.05 = p(e^{sign}|x_i, n_i^{add}, n_i^{ini}, H_0)/.05$. Thus, $ratio_\alpha$ is above 1 if the α-error of a study—conventionally set at .05—increases due to motivated conditional data collection.

Results

The probability of a conditional false positive increases when the upper bound x_i for p^{crit} and the number of additional participants run decreases (Figure 1, A, left). This means that a researcher who observes (e.g.) $p^{crit} = .055$ risks a 40% chance for a false positive (i.e., H_0 is true but $p < .05$) when s/he practices to run two additional participants in the case of an initially observed p-value between .05 and .06. More initial participants generally amplify the rate of false positives (Figure 1, A, middle and right).[2] For an increasing number of initial participants run, the probability of a false positive depends less on the specific number of additional participants run when the upper bound x_i for the range of critical p-values is low.

In all conditions, the α-error conventionally set at .05 increases and thus $ratio_\alpha$ is above 1 (Figure 1, B). Contrary to the probability of a false positive, the ratio decreases when x_i and the number of additional participants decrease (Figure 1, B, left) or the number of initial participants run increases (Figure 1, B, middle and right). This is due to a lower chance of observing a critical p-value close to .05 (i.e., e^{crit}) for a decreasing x_i and of observing an even lower p-value when the initial number of participants was already high.

To summarize, results show that the probability for a false positive is highest when a high number of initial participants produces an initial p-value close to the significance level and a low number of additional participants is run. On the contrary, the increase of the α-error is lowest for these conditions because the event of a critical p-value is, a-priori, lower for these conditions.

[1] An alternative stopping rule for motivated conditional data collection is running additional participants as long as the observed p-value decreases with each subsequent significance test. This behavior further increases the probability of a false positive. Results from the simulations reported here can therefore be considered as a lower benchmark.

[2] I checked the stability of this trend within an upper range of $n_i^{add} = 100$.

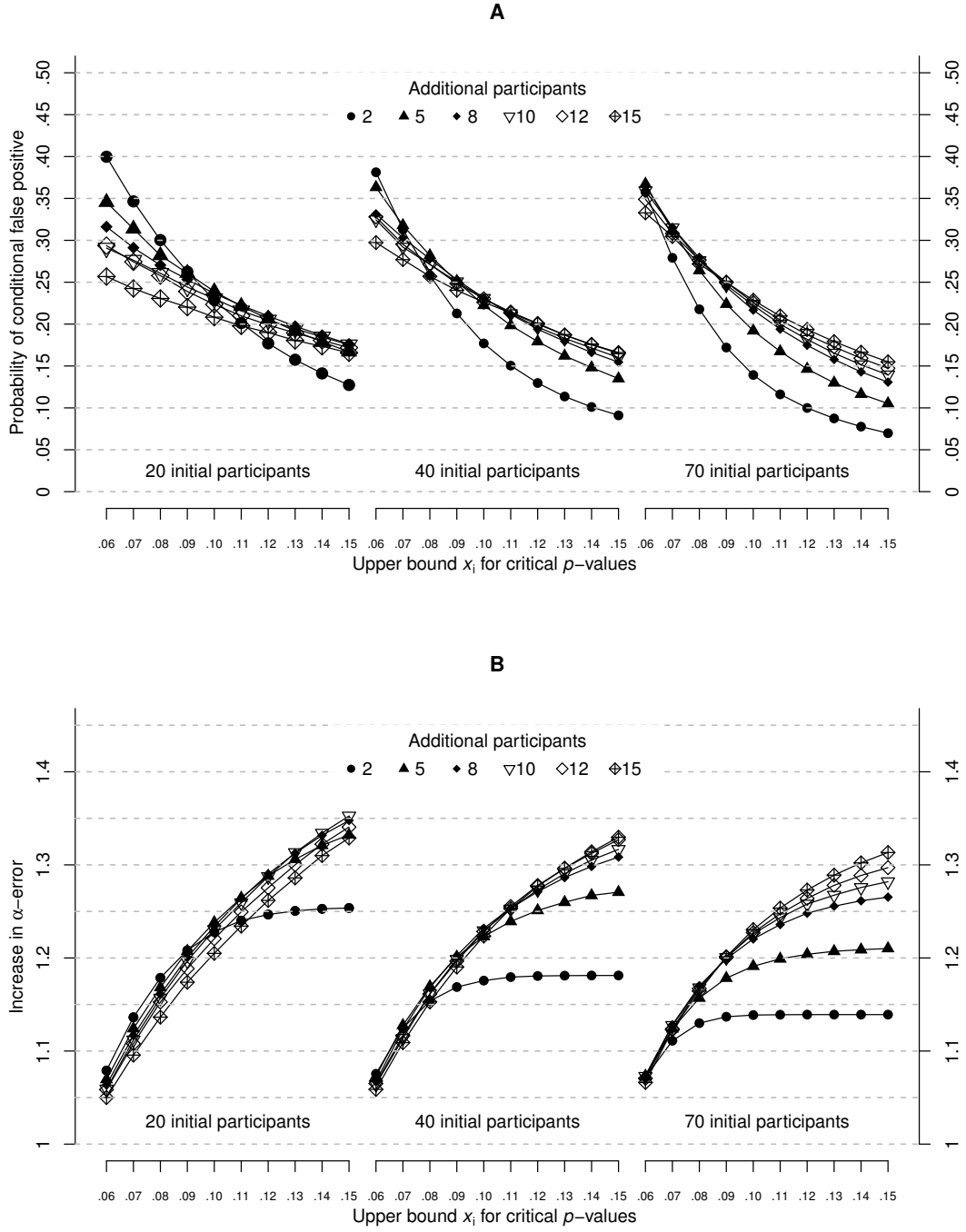

Figure 1. Probability of a conditional false positive (panel A) and increase in α-error due to motivated conditional data collection (panel B) dependent on the initial number of participants n_i^{ini}, the number of additional participants n_i^{add}, and the upper bound x_i for the range of critical p-values p^{crit}.

Does a p-value close to .05 signal H_1 being true?

One may believe that it is more likely that there is an effect (i.e., H_1 is true) versus no effect (i.e., H_0 is true) when observing a p-value close to the desired significance level (i.e., e^{crit} is given), or expressed in posterior odds: $Odds_i = \frac{p(H_1|e^{crit},x_i,n_i^{ini})}{p(H_0|e^{crit},x_i,n_i^{ini})} > 1$. If e^{crit} signals H_1 being true, motivated conditional data collection may appear justified in order to minimize the probability of an omission error in the detection of an effect also known as the β-error of a study (Fiedler, Kutzner, & Krüger, 2012). According to Bayes' rule, $Odds_i$ can be calculated by multiplying the ratio between the priors of the hypotheses $ratio(priors) = \frac{p(H_1)}{p(H_0)}$ with the likelihood ratio $ratio(L)_i = \frac{p(e^{crit}|x_i,n_i^{ini},H_1)}{p(e^{crit}|x_i,n_i^{ini},H_0)}$ of the critical event e^{crit} of an initial p-value for x_i under H_1 and H_0, or: $Odds_i = ratio(priors) \times ratio(L)_i$. For (e.g.) $x_i = .07$, a test power of $(1-\beta)_i = \{.80, .50, .95\}$, an α-error = .05, and a medium effect size of Cohen's $d = .50$ for H_1 and therefore $n_i^{ini} = \{27, 21, 45\}$ (Faul, Erdfelder, Buchner, & Lang, 2009), a simulation based on 500,000 studies for each condition shows that $ratio(L)_i$ equals $\{2.28, 2.92, 0.78\}$. Thus, it is about twice or three times as likely to observe e^{crit} under H_1 vs. H_0 for the test power of .80 recommended for studies or .50 usually acquired in studies (Sedelmeier & Gigerenzer, 1989) but less likely for well-designed studies with α-error = β-error = .05. Taking into account the incentive to test counter-intuitive hypotheses because they are more likely published (Ioannidis, 2005) and thus assuming a generally lower prior probability of H_1 (i.e., $ratio(priors) < 1$), it is premature to infer H_1 being true (i.e., $Odds_i \gg 1$) after observing e^{crit} and to feel justified to run additional participants.

Scenario 2: Continuous motivation for conditional data collection

Researchers may not always run additional participants after observing a p-value below a specific threshold; they most likely do not even possess a fixed threshold at all. It is more likely that the motivation for running additional participants increases continuously for a decreasing p-value, such that it is more likely but not certain for a researcher to collect additional data when observing a p-value of (e.g.) .07 than .20. In the following scenario, I therefore simulate a researcher with a continuous motivation for running additional participants to make the scenario more realistic and to test for the stability of the results from the first study.

Method

I simulated six types of researchers with different motivation functions f^{motiv} that translate an initially observed p-value p^{ini} into a probability for running additional participants (Figure 2, panel C). I thereby simulate researchers who differ in their sensitivity to the distance from the initially observed p-value to the desired significance level for running additional participants. The function f^{motiv} is based on the Beta density function that produces densities above zero for values between 0 and 1 and whose shape is dependent on two parameters. I fixed the first parameter of the Beta density function at 1 and varied the second parameter with $s_i \in S = \{3, 5, 7, 10, 25, 50\}$, resulting in right-skewed functions that vary in steepness. To receive probabilities instead of densities from p^{ini} and to set the maximum probability of 1 for motivated conditional data collection given $p^{ini} = .05$ and a minimum probability of 0 given $p^{ini} = 1$, I divided the Beta density function with the density of the Beta function for s_i at a value of .05, or: $f_{s_i}^{motiv}(p^{ini}) = B(p^{ini}|1, s_i)/B(.05|1, s_i)$. Higher values in s_i result in steeper probability curves (Figure 2, panel C). A researcher with an underlying motivation function with (e.g.) $s_i = 50$ has a high probability for running additional participants only for initial p-values close to the significance level of .05, whereas a researcher with $s_i = 3$ is less sensitive to the distance from the initial p-value to the significance level and also tends to run additional participants for higher initial p-values. Except for the motivation function and a fixed number of 20 initial participants run, methods and conditions are identical to the first study.

Results

For a high sensitivity (i.e., $s_i = 50$ and $s_i = 25$) to the distance from the initial p-value to the desired significance level, the pattern from study 1 can be replicated: a low number of additional participants leads to the highest probability for a conditional false positive (i.e., .28; see Figure 2, panel A). Also in accordance with the results from the first study, the ratio in the α-error increases up to 1.45 for a decreasing sensitivity (i.e., $s_i = 3$) and an increasing number of additional participants (Figure 2, panel B). In summary, scenario 2 replicates the findings from scenario 1 and extends results from discrete to the more realistic scenario of a continuous motivation for running additional participants.

Discussion

I have shown, for types of researchers that vary in their motivation to collect further data depending on the initially observed p-value, that the chance for a conditional false positive can increase up to 40% and the conventional α-error of .05 can increase by a factor of 1.45 due to motivated conditional data collection. Given that an estimated 72% of researchers practiced this behavior in the past, this means that for every 1,000 false positives due to the conventional α-error of .05 in null-hypothesis significance testing, an additional number up to 324 false positives occur because of motivated conditional data collection.[3] Finally, I have also identified a

[3] $1.45 \times .72 + 1 \times .28 = 1.324$

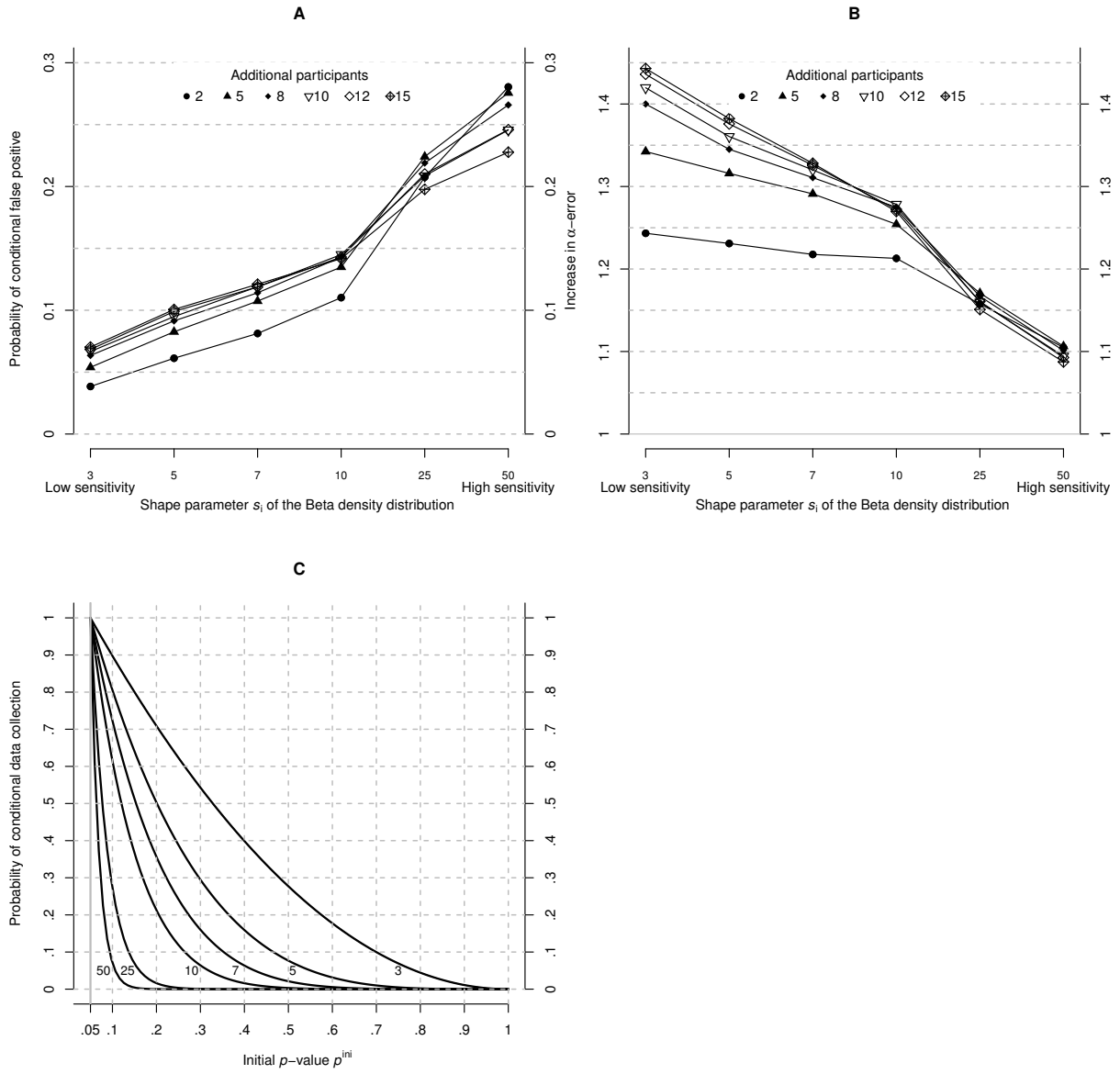

Figure 2. Probability of a conditional false positive (panel A) and increase in α-error due to motivated conditional data collection (panel B) dependent on the initial number of participants $n_i^{ini} = 20$, the number of additional participants n_i^{add}, and the underlying motivation function $f_{s_i}^{motiv}$ for conditional data collection with higher s_i—as indicated by a number close to each curve—resulting in researchers who are more sensitive to the distance from the initial p-value to the desired significance level (panel C).

set of false (as demonstrated by simulations) beliefs that may be partially the reason for the high prevalence of motivated conditional data collection.

In an editorial in *JESP*, Cooper (2012) argues that "the misdeeds of some of our members are indictments of those individuals and a betrayal of our field" and not "indictments of our field". Although this is true for extreme behavior like excluding outliers or falsifying data, our field should take responsibility of the naïve misdeeds of less extreme but common and thus likely more damaging questionable research behavior like motivated conditional data collection. Targeting false beliefs in teaching interventions might be a first step in this direction.

References

Cooper, J. (2012). On fraud, deceit and ethics. *Journal of Experimental Social Psychology, 49*, 314

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

Feller, W. (1940). Statistical aspects of ESP. *Journal of Parapsychology, 4*, 271–298.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The Long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 6*, 661–669.

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

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science, 6*, 639–642.

Gadbury, G. L. (2012). Inappropriate fiddling with statistical analyses to obtain a desirable *p*-value: Tests to detect it presence in published literature. *PLOS ONE, 7*, 1–9.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*, 0696–0701.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 6*, 524–532.

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

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.

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

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