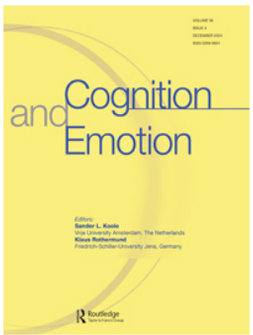

Cognition and Emotion

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/pcem20

Credibility of results in emotion science: a *Z*-curve analysis of results in the journals *Cognition & Emotion* and *Emotion*

Maria D. Soto & Ulrich Schimmack

To cite this article: Maria D. Soto & Ulrich Schimmack (20 Dec 2024): Credibility of results in emotion science: a *Z*-curve analysis of results in the journals *Cognition & Emotion* and *Emotion*, Cognition and Emotion, DOI: 10.1080/02699931.2024.2443016

To link to this article: https://doi.org/10.1080/02699931.2024.2443016

COGNITION AND EMOTION
https://doi.org/10.1080/02699931.2024.2443016

Routledge
Taylor & Francis Group

Credibility of results in emotion science: a *Z*-curve analysis of results in the journals *Cognition & Emotion* and *Emotion*

Maria D. Soto and Ulrich Schimmack

Department of Psychology, University of Toronto Mississauga, Mississauga, Canada

ABSTRACT

Failed replication attempts have raised concerns over the prevalence of publication bias and false positive results in the psychological literature. Using a sample of 65,970 test statistics from *Cognition & Emotion* and *Emotion*, this article assesses the credibility of results in emotional research. All test statistics were converted to *z*-scores and analysed with *Z*-curve. A *Z*-curve analysis provides information about the amount of selection bias, the expected replication rate and the false positive risk. Lastly, *Z*-curve is used to determine an alpha level that lessens the false positive risk without unnecessary loss of power. The results show evidence of selection bias in emotional research, but trend analyses showed a decrease over time. Based on the *z*-curve estimates, we predict a 15% and 70% success rate in replication studies. Therefore, replication studies should increase sample sizes to avoid type-II errors. The risk of false positives with the traditional alpha level of 5% is between 5% and 33%. Lowering alpha to 1% is sufficient to reduce the false positive risk to less than 5%. In sum, our findings may alleviate concerns about high false positive rates among emotional researchers. However, selection bias and low power remain challenges to be addressed.

ARTICLE HISTORY
Received 26 April 2024
Revised 30 November 2024
Accepted 11 December 2024

KEYWORDS
Meta-analysis; statistical power; publication bias; *Z*-curve; false positive risk

Emotion research reemerged from the dark ages of behaviourism in the 1980s. During these early years, emotion research was published in a wide range of journals. Since 1987, *Cognition & Emotion* was established as a journal dedicated to the study of emotion. In 2001, the American Psychological Association created the journal *Emotion* for the same purpose. Over the past two decades, these two journals have published hundreds of articles that report the results of empirical studies on emotions. Given the high costs of experimental designs, many studies have modest sample sizes. This raises several concerns. First, effect size estimates in small samples are imprecise, and point estimates are inflated when results are selected for significance. Second, significant results can be difficult to replicate because studies have only modest power. Finally, a large portion of statistically significant results may be false positive results (i.e. the population effect size is close to zero or the sign of the effect is in the opposite direction to the reported result).

In the past decade, psychology has been shaken by fraud scandals and replication failures of textbook findings. A reproducibility project replicated 100 studies and only 36% of replication attempts reproduced a significant result (Open Science Collaboration, 2015). While the results for cognitive psychology were slightly better (50%), the results for social psychology were worse (25%). In emotion research, the textbook finding that manipulations of facial muscles with the pen paradigm change emotional experiences failed to replicate in two large replication studies (Coles et al., 2022; Wagenmakers et al., 2016), providing strong evidence that the original results reported by

CONTACT Maria D. Soto maria.soto@mail.utoronto.ca Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Road, Mississauga, ON L5L 1C6, Canada

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02699931.2024.2443016.

Strack et al. (1988) that have been cited over 2000 times were false positive results.

These replication failures in the Open Science Reproducibility project have raised concerns that many, if not most, published results might be false positives (Ioannidis, 2005; Simmons et al., 2014). One problem with thisstudy is that it is unclear whether these results can be generalised to other areas like emotion research. Another problem is that the results are limited to articles in the year 2008. In response to concerns about the credibility of results in psychology journals, psychological journals have embraced open science practices such as data sharing and preregistration of analysis plans. However, it is not clear how much these practices have improved the credibility of published results.

This article aims to provide a comprehensive assessment of the credibility of results in *Cognition & Emotion* and *Emotion*, using *Z*-curve (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020; Schimmack, 2020; Schimmack & Bartoš, 2023). *Z*-curve relies on the strength of evidence against the standard null-hypothesis of no effect in either direction. All statistical tests are converted into two-sided *p*-values, which in turn are converted into absolute *z*-scores. Larger *z*-scores are less compatible with the null-hypothesis. *Z*-curve uses this information to estimate two parameters that can be used to evaluate the credibility of published results, namely the Expected Discovery Rate (EDR) and the Expected Replication Rate (ERR).

The expected discovery rate, selection bias and false discovery risk

Z-curve is a selection model that assumes the selection of results into the literature is a function of a study's power. A true null-result has only a 5% probability of being published. A study with 80% power has an 80% probability of being published. This mixture of powers produces a distribution of *z*-scores that can be used to estimate the mean power of studies before selection for significance (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020). As mean power determines the percentage of significant results, Bartoš and Schimmack (2022) call this estimate the Expected Discovery Rate (EDR) because the term discovery rate is used in the statistical literature to refer to the percentage of significant results (Soric, 1989). In short, a set of studies with a mean power of 50% is expected to produce 50% significant results.

The *z*-curve estimate of the EDR can be used to quantify the amount of selection bias in emotion journals. Selection bias – also called publication bias – refers to the preferential publication of significant results over non-significant results. Concerns about selection bias in psychology journals were first raised by Sterling (1959) who found that psychology journals report over 90% statistically significant results. This finding has been replicated decade after decade (Fanelli, 2010; Motyl et al., 2017; Sterling et al., 1995). In the past 10 years, concerns have been raised that researchers are using questionable research practices to increase their chances of getting a publishable significant result (John et al., 2012). These practices increase the risk of publishing a false positive result (Simmons et al., 2011). The low replication rate in the Reproducibility Project raised concerns that many of the replication failures revealed false positive results in original studies that were obtained with QRPs.

Z-curve's estimate of the EDR makes it possible to quantify the amount of selection bias in the literature by comparing the Observed Discovery Rate (ODR) with the EDR. The ODR is simply the percentage of statistical results with *p*-values below the conventional significance criterion, alpha = .05. Without selection bias, the ODR should match the EDR (Brunner & Schimmack, 2020). However, the ODR can be much higher than the EDR, if results are selected for significance. For example, Schimmack (2020) used results from social psychology journals hand-coded by Motyl et al. (2017) and found an ODR of 90%, but the EDR was only 19%. The large discrepancy of 71 percentage points reveals selection bias in social psychology. Another example comes from abstracts of medical articles that reported clinical trials (Schimmack & Bartoš, 2023). Whereas the ODR was 69%, the EDR was 29%. These results can be used as a comparison standard for emotion research.

The EDR also provides valuable information about the false discovery risk. In statistics, the false discovery rate is defined as the percentage of significant results that were obtained when the null-hypothesis is true. For example, an FDR of 20% implies that the null-hypothesis is true for 1 out of 5 statistically significant results. Speculations about false discovery rates vary widely based on untested assumptions about power and the number of false hypotheses that are being tested. One view is that the null-hypothesis is rarely true and that the risk of a false positive result is low (Cohen, 1994). Another view is that researchers are

much more likely to test false hypotheses than true hypotheses and that the false discovery rate could be over 50% (Ioannidis, 2005).

Z-curve does not require untestable assumptions and can provide empirical estimates of the false discovery risk. Using a formula from Soric (1989), it is possible to compute the maximum false discovery rate based on the discovery rate. So far, estimates of the FDR required access to all statistical tests. This made the formula useless when selection bias is present. However, *Z*-curve's estimate of the EDR when selection bias is present can be used to estimate the false discovery risk using Soric's formula. For example, with an inflated observed discovery rate of 90%, the implied FDR for social psychology would be only 1%. However, with an EDR of 19%, the FDR for social psychology is 22%, that is 1 out of 4–5 results could be false positives.

Bartoš and Schimmack (2022) call Soric's maximum False Discovery Rate, the False Discovery Risk (FDR). The reason is that it is impossible to estimate the actual rate of false positive results, but there is a risk that up to 22% of results in social psychology could be false positives. While this estimate may seem high, the results refute claims that most published results in psychology are false (Ioannidis, 2005).

Aside from estimating the FDR for the conventional significance criterion of alpha = .05, *Z*-curve can also be used to control the risk of false discoveries by adjusting alpha. Statisticians have argued that the standard significance criterion, $p < .05$, contributes to the replication failures in the reproducibility project because it is too easy to obtain significant results with this criterion, especially when QRPs are used (Simmons et al., 2011). Benjamin et al. (2018) proposed to lower alpha to .005. However, this suggestion increases the risk of false negative results, especially in research areas that cannot easily increase sample sizes to compensate for the loss in statistical power. Moreover, this suggestion was based on hypothetical assumptions that may not match the research practices of emotion researchers. *Z*-curve makes it possible to adjust alpha enough to reduce the false positive risk without unnecessary loss of power. For example, Schimmack and Bartoš (2023) found that in medical journals the FDR of 14% with alpha = .05 could be reduced to an FDR of 4% with alpha = .01. Our results for emotion journals can be used to adjust alpha to reduce the FDR in emotion research to a reasonable level without an unnecessary loss of power.

In sum, a *Z*-curve analysis of significant results in emotion journals produces an estimate of the Expected Discovery Rate (EDR). The EDR can be used to quantify the amount of selection bias in the emotion literature. It can also be used to estimate the false discovery risk. Finally, *Z*-curve can be used to determine an alpha level that produces an acceptable false discovery risk.

The expected replication rate

The expected replication rate is the power of studies that produced a significant result and were published. In theory, the ERR makes it possible to estimate the percentage of significant results in replication studies because mean power determines the success rate in a set of exact replication studies with the same sample sizes as the original studies (Brunner & Schimmack, 2020). However, is often difficult to conduct exact replication studies. This may explain why the *Z*-curve estimate of the ERR for the Reproducibility Project was higher than the actual rate of 36%. Bartoš and Schimmack (2022) therefore argued that the ERR is an optimistic estimate of the maximum significant results that can be expected, while the EDR provides a minimum. Estimates of the ERR are useful for sample size considerations of replication studies. To avoid replication failures of true hypotheses, it is necessary to take the power of original studies into account.

Method

Extraction of test statistics

The complete repertoire of published articles by *Cognition & Emotion*, from 1987 to 2023, and *Emotion*, from 2001 to 2023, were collected as PDF files for the project. The reported test statistics (F, t, χ^2, z, 95% CI) of each study were systematically extracted from each PDF file through a custom R-code. Additionally, we extracted 95% confidence intervals of odds ratios and regression coefficients.

The chi-square test statistics and the 95% confidence intervals had to meet certain conditions to be included in the analysis. The code identified 396 chi-square tests with degrees of freedom over 6 in *Emotion* (22.64%) ranging between 7 and 2484, and 253 in *Cognition & Emotion* (16.71%) ranging between 7 and 1861. Most chi-squares above 6 were larger than 10 (12.42%) in *Cognition & Emotion,*

Table 1. Total *Z*-scores per test statistic.

	Cognition & Emotion, $N = 30{,}513$[a]	*Emotion*, $N = 35{,}457$[a]
Statistic type		
F	18,977 (62%)	18,808 (53%)
t	7827 (26%)	10,053 (28%)
95% CI	1053 (3.5%)	2785 (7.9%)
χ^2	1261 (4.1%)	1353 (3.8%)
PESE	712 (2.3%)	1412 (4.0%)
z	683 (2.2%)	1046 (3.0%)

[a]*n* (%).

while only 4.29% were between 6 and 10. Similarly, in *Emotion,* 17.44% of the chi-squares above 6 were larger than 10, and 5.20% between 6 and 10. Only chi-square tests with degrees of freedom between 1 and 6 were extracted. The reasoning behind this condition is to exclude chi-square tests performed for model testing in structural equation modelling articles and that in these tests a strong rejection of the null-hypothesis reveals poor model fit rather than support for a theoretical prediction.

Next, confidence intervals were excluded when they were reported in addition to test statistics to avoid counting the same result twice.

Lastly, test statistics from meta-analyses were excluded because *Z*-curve relies on individual-level test statistics, and it is not guaranteed that the meta-analysis reported all statistics for every study included. Additionally, meta-analyses would have introduced test statistics that were not originally published by the journals of interest.

The code relies on the *pdftools* R package (Ooms, 2024) to render all textboxes from the PDF files into processable character strings. Working with various journals presents a challenge to ensure that all or at least most notation formats are accounted for to ensure the maximum extraction of test statistics. Consequently, the r-code is designed to accommodate various notation formats, and it has been tested against multiple journals across disciplines. Furthermore, the original r-script was fine-tuned to handle

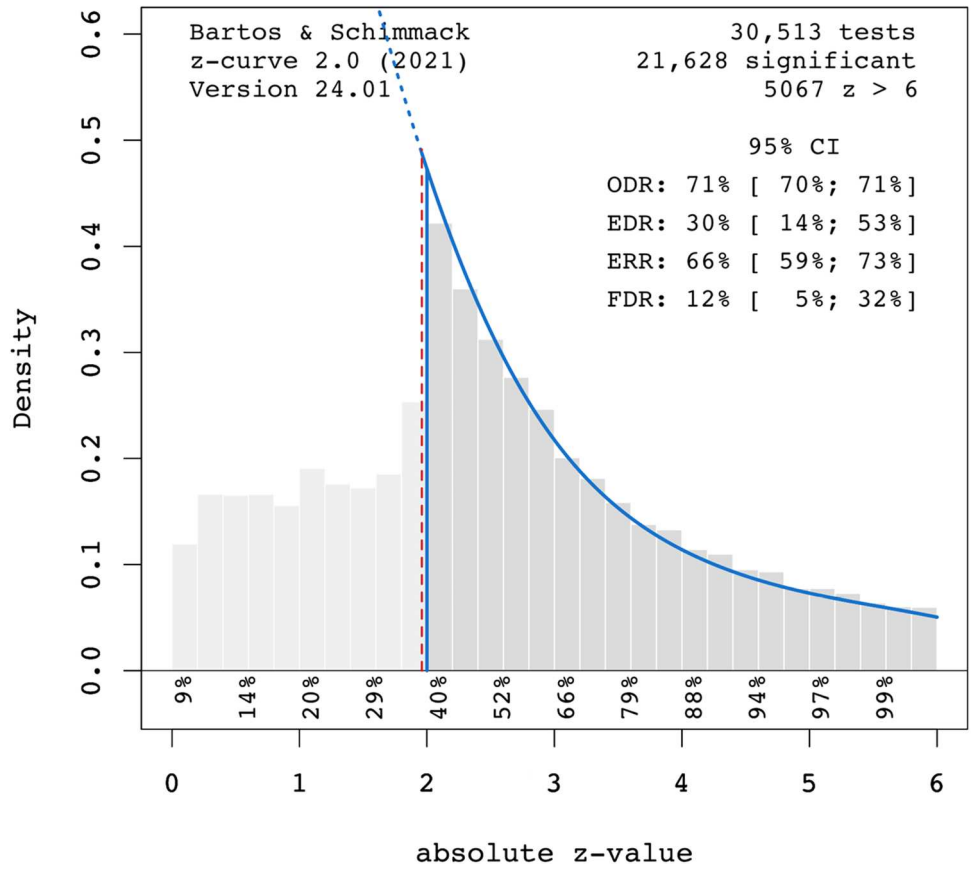

Figure 1. *Z*-curve plot for *Cognition & Emotion*.

Note: ODR = Observed Discovery Rate, EDR = Expected Discovery Rate, ERR = Expected replication rate, FDR = False Discovery Risk.

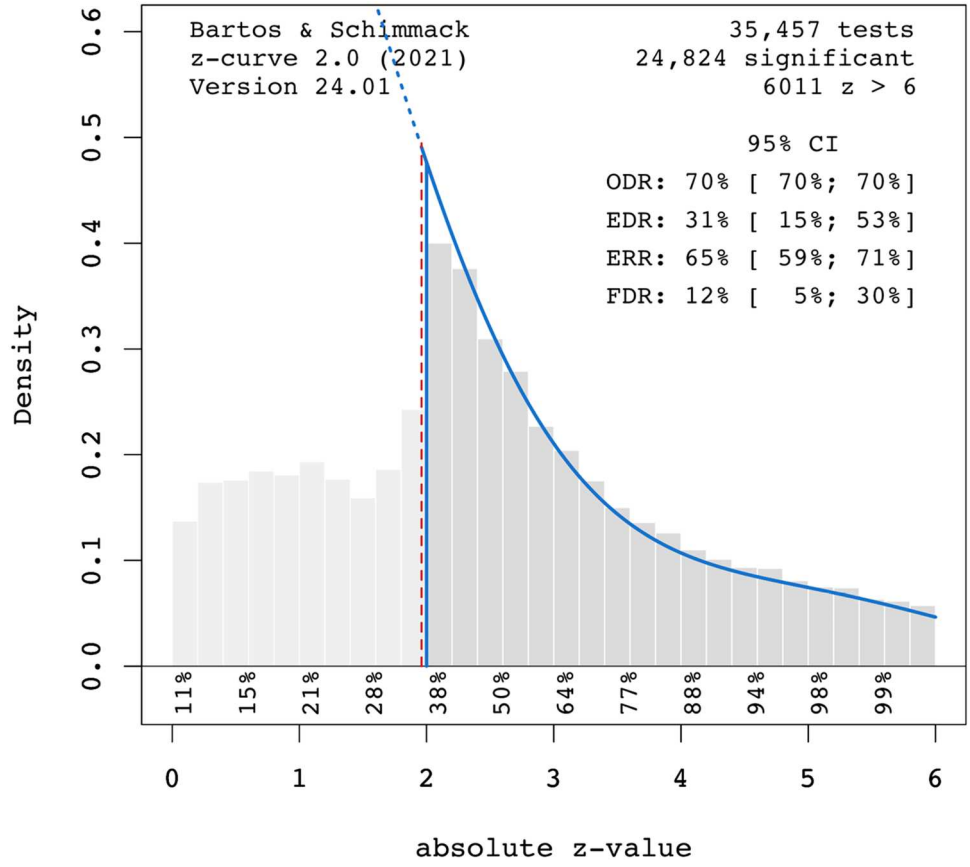

Figure 2. *Z*-curve plot for Emotion.

Note: ODR = Observed Discovery Rate, EDR = Expected Discovery Rate, ERR = Expected replication rate, FDR = False Discovery Risk.

specific notation patterns present in the journals *Cognition & Emotion* and *Emotion*. However, it is still possible that some errors made it into the final sample.

Although the r-code proficiently extracts test statistics from text paragraphs, it cannot extract those reported within tables or figures. This limitation remains unless the statistic is reported in a predetermined notation format, as demonstrated by tables reporting a series of *F* statistics as "*F* (2, 145) = 3.13, $p < .05$", which includes all necessary components for extraction. Additionally, the automated extraction process cannot distinguish between focal and non-focal results. Following extraction, the test statistics were converted into absolute *z*-scores.

Statistical analysis

Utilising the *z*-curve package in R (Bartoš & Schimmack, 2022) the objective was to assess the credibility of results in *Cognition & Emotion* and *Emotion*. To account for clustering, we utilised the "*b*" method from the *zcurve_clustered* function as it samples a single test statistic from each article for model fitting. *Z*-curve uses the expectation maximisation (EM) algorithm to fit the distribution of the observed statistically significant results for *z*-scores between 1.96 and 6 (Bartoš & Schimmack, 2022). Values above 6 are treated as tests with 100% power. *Z*-curve estimates the optimal weights for each component out of seven components ($z = 0{:}6$) to fit the observed distribution of the significant *z*-scores. Following model fitting, *Z*-curve extrapolates the full distribution, thereby estimating the shape of the distribution of the statistically non-significant results (Bartoš & Schimmack, 2022). The weights are used to compute the Expected Discovery Rate (EDR) and the Expected Replication Rate (ERR). The Observed Discovery Rate (ODR) is simply the percentage of significant results, $p < .05$. The False Discovery Rate (FDR) is a simple transformation of the EDR using Soric's (1989) formula.

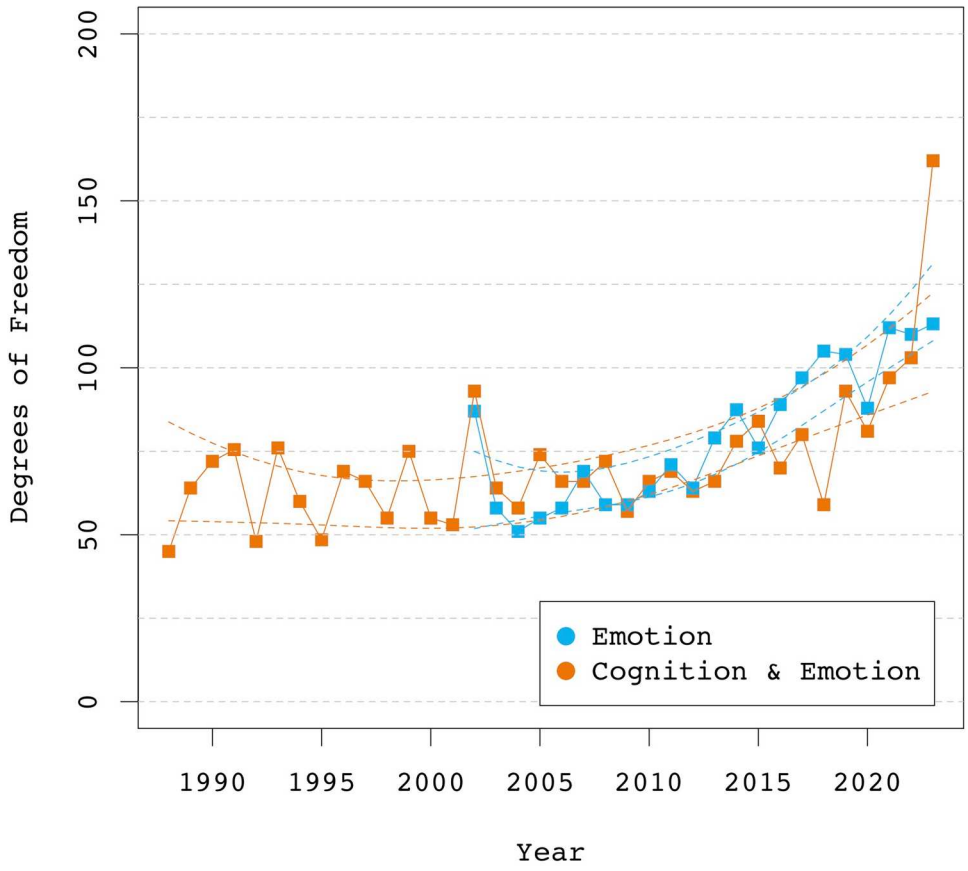

Figure 3. Trends of within-group degrees of freedom from *Cognition & Emotion and Emotion* from 1988 to 2023.

Selection for significance

The ODR is the percentage of observed statistically significant results. Meanwhile, the EDR is the expected discovery rate based on the mean power of studies before selection for significance. Comparing the proportion of observed significant results (ODR) to the (EDR) quantifies the amount of selection bias present. The higher the difference is, the more effect size estimates of studies before selection for significance are inflated.

Expected replication rate

The ERR is the mean power estimate after selection for significance. The ERR is higher than the EDR, mean power before selection for significance, because selection for significance favours studies with high power. A study with 80% power has a higher chance of being published than a study with 20% power. Given the relationship between power and replicability, this estimate predicts the anticipated frequency in which statistically significant would replicate with the same sample sizes as the original studies.

False discovery risk

Z-curve can estimate the False Discovery Risk (FDR) based on Soric's formula that determines the maximum false discovery rate compatible with the discovery rate. When selection bias is present, the EDR estimate is used to estimate FDR (Bartoš & Schimmack, 2022).

Time trends

The *Z*-curve analyses of all data were followed up by separate analyses for each publication year. These annual estimates were regressed on a linear and quadratic predictor of publication year to examine time trends. A quadratic term was included as a predictor to test the hypothesis that EDR and ERR estimates remained constant before 2011 and increased only in the past decade in response to the replication crisis.

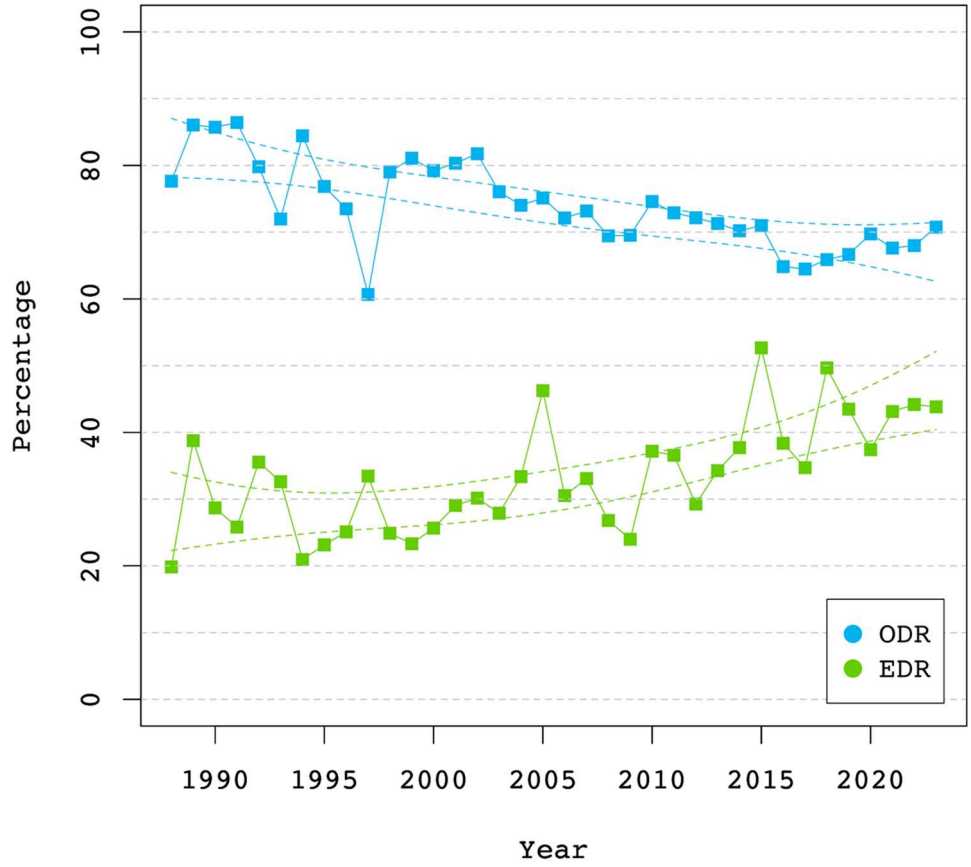

Figure 4. The observed and expected discovery rate of *Cognition & Emotion*.
Note: ODR = Observed Discovery Rate, EDR = Expected Discovery Rate.

Results

Description of sample characteristics

We downloaded 3831 articles from *Cognition & Emotion* (1987–2023) and 2323 articles from *Emotion* (2001–2023). Articles such as editorials, review papers, and meta-analyses were excluded. Of the collected files, 2028 articles from *Cognition & Emotion* and 1955 articles from *Emotion* included at least one statistical result that could be used for the *z*-curve analyses.

Not enough data was available to perform an annual *Z*-curve for 1987 from *Cognition & Emotion* and for 2001 from *Emotion*. The statistics from each of these were joined with the following year, meaning 1988 contains 2 articles published in *Cognition & Emotion* in 1987 and 2002 contains 10 articles published in *Emotion* in 2001. Additionally, test statistics with sample sizes below 30 participants were excluded because the conversion of test statistics (*t*, *F*) into *z*-scores does not approximate the standard normal distribution (Schimmack, 2024). However, additional *Z*-curve plots performed on the complete sample and other exclusions of interest can be found in the supplementary materials (https://osf.io/42vxd/), these analyses indicate the present results are robust with similar parameter estimates. In total, 5796 (15.91%) of extractable test statistics were excluded from the *Cognition & Emotion* sample and 6486 (15.46%) from the *Emotion* sample. The present results were run on a set of 1902 articles from *Cognition & Emotion* and 1953 from *Emotion*.

The total test statistics extracted were 30,513 for *Cognition & Emotion* and 35,457 for *Emotion*. Most of the test statistics were *F* and *t*-tests (Table 1). The median degree of freedom for *F*-tests and *t*-tests was 67.5, ranging from 45 to 162 for *Cognition & Emotion* and 77, ranging from 51 to 106.113.16 for *Emotion*.

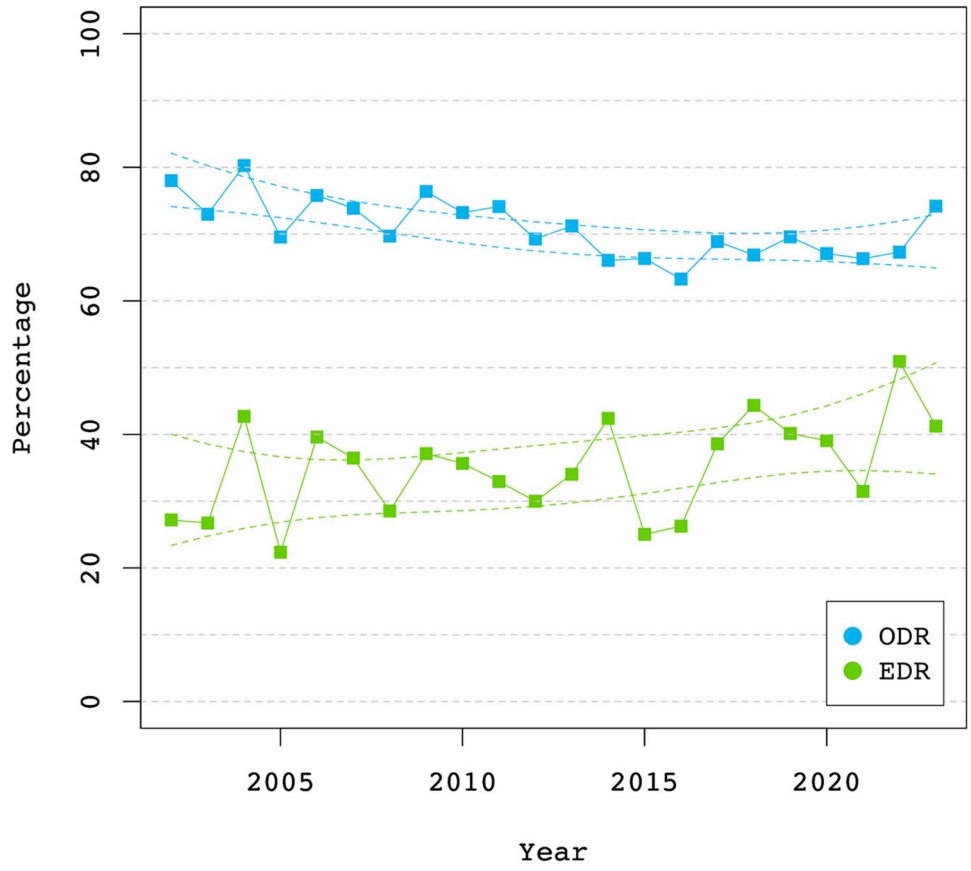

Figure 5. The observed and expected discovery rate of *Emotion*.
Note: ODR = Observed Discovery Rate, EDR = Expected Discovery Rate.

Z-curve estimates

Figure 1 and Figure 2 show the histogram of *z*-values for *Cognition & Emotion* and *Emotion*, respectively. Both journals show very similar Observed Discovery Rates (ODR), C&E: 71% (95% CI [70%; 71%]); Emo: 70% (95% CI [70%; 70%]). Both journals have notably lower expected discovery rates, and their 95% confidence intervals do not include the ODR estimates, C&E: 30% (95% CI [14%; 53%]); Emo: 31% (95% CI [15%; 53%]). Thus, there is clear evidence of selection bias in both journals. This conclusion is also consistent with visual inspection of the Figures that show a sharp drop in observed *z*-scores just below the value for statistical significance (1.96).

The expected replication rates (ERR) were almost identical, C&E: 66% (95% CI [59%; 73%]); Emo: 65% (95% CI [59%; 71%]). The estimates suggest that replication studies with the same sample size should replicate more often than not. However, it is important to note that these are optimistic estimates of actual replication rates and that the true replication rate is likely to be lower due to problems in conducting exact replications. Last, the false discovery risks were also similar and not significantly different from each other, C&E: 12% (95% CI [5%; 32%]); Emo: 12% (95% CI [5%; 30%]). These estimates are similar to those for clinical trials in medical journals, 14% (Schimmack & Bartoš, 2023) and much lower than one would expect based on concerns that most published results are false (Ioannidis, 2005). Our estimate of the ERR implies that we expect about 40% replication failures, while our estimate of the FDR is only 12%. Thus, replication failures should not be considered evidence of a false discovery in original studies, unless the replication study had a much larger sample size. It is important to note that although the FDR in both journals is 12%, their upper confidence interval is around 30%. This level is unacceptably high and suggests that a lower alpha level is needed to maintain a reasonable false

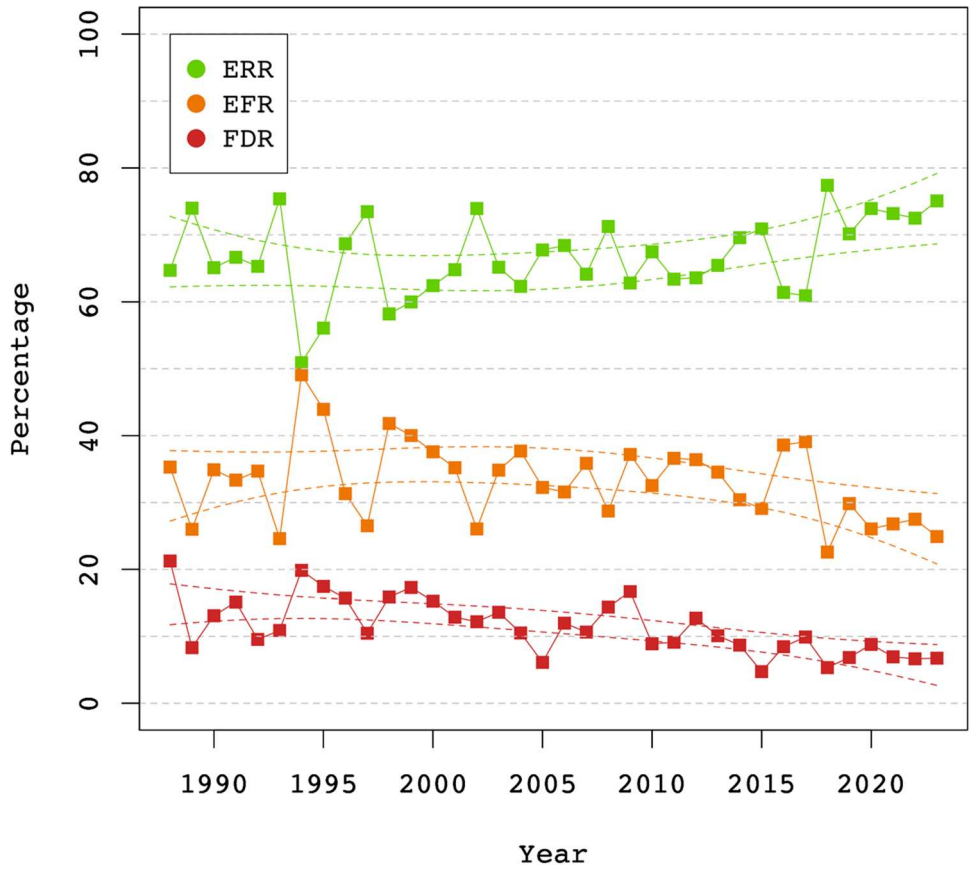

Figure 6. The false positive risk and replicability of *Cognition & Emotion*.
Note: ERR = Expected Replication Rate, EFR = Expected Replication Failure Rate (1 − ERR), FDR = False Discovery Risk.

positive risk. Furthermore, the FDR refers to findings with a statistically significant result used to claim that the null hypothesis is false, it does not exclude statistically significant results with a trivial effect size. Thus, the estimates do not measure the proportion of statistically significant results that may lack practical significance.

Changes over time

Degrees of freedom

Cognition & Emotion and *Emotion* both showed significant linear increases over time to the degrees of freedom of *F*-tests and *t*-tests, $b = 1.11$, SE = 0.25; $p < 0.0001$ and $b = 2.68$, SE = 0.32; $p < 0.0001$, respectively. Additionally, both journals showed a significant quadratic trend, Emo: $b = 0.17$, SE = 0.06; $p = 0.007$, and C&E: $b = 0.08$, SE = 0.03; $p = 0.003$. Figure 3 offers a visual overview of the trends present in both journals. Thus, there is some evidence that the replication crisis has produced an increase in sample sizes. However, sample sizes were already on an upward trend.

Observed and expected discovery rates

As seen in Figures 4 and 5, both journals showed similar decreases in the ODR over time, C&E: $b = -0.45$, SE = 0.07; $p < 0.0001$; Emo: $b = -0.44$, SE = 0.11; $p = 0.001$. No significant quadratic trends were observed for *Cognition & Emotion*, $b = 0.004$, SE = 0.01, $p = 0.531$, nor *Emotion*, $b = 0.04$, SE = 0.02, $p = 0.070$. This finding suggests that researchers are reporting non-significant results more often over time, but not in response to the replication crisis. Importantly, it is not clear whether the reporting of non-significant results also increased for focal hypothesis tests or whether it just became more common to report statistical results for non-significant results rather than not reporting these results or reporting them without quantitative information (e.g. $F < 1$ or ns).

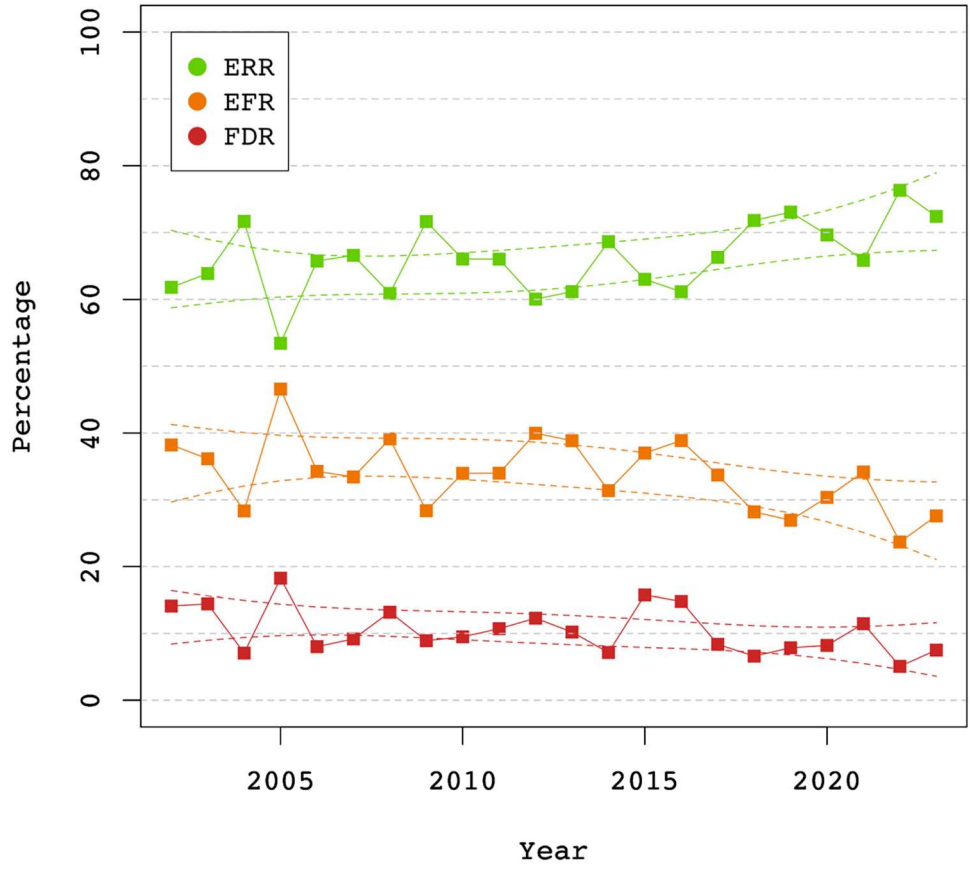

Figure 7. The false positive risk and replicability of *Emotion*.
Note: ERR = Expected Replication Rate, EFR = Expected Replication Failure Rate (1 – ERR), FDR = False Discovery Risk.

Both journals showed an increase in the EDR over time (Figures 4 and 5), which is consistent with the increase in the degrees of freedom (sample sizes), C&E: $b = 0.52$, SE = 0.10, $p < 0.0001$; Emo: $b = 0.51$, SE = 0.23; $p = 0.038$. The non-linear trends were not significant, C&E: $b = 0.02$, SE = 0.01, $p = 0.074$; Emo: $b = 0.03$, SE = 0.04, $p = 0.505$. In combination, these results suggest that selection bias has decreased over time, although it is still present in the latest years.

Expected replicability rates and false discovery risks

Consistent with the higher degrees of freedom (sample sizes), the ERR of both journals increased over time, C&E: $b = 0.18$, SE = 0.09, $p = 0.044$; Emo: $b = 0.41$, SE = 0.16, $p = 0.018$ (Figure 6 and 7). Additionally, *Cognition & Emotion* showed a significant non-linear trend, $b = 0.02$, SE = 0.01, $p = 0.047$. No significant non-linear trend was observed for *Emotion*, $b = 0.04$, SE = 0.03, $p = 0.201$. The findings suggest the replication crisis may have prompted changes that improved the replicability of findings in *Cognition & Emotion*. Figures 6 and 7 also show the Expected Replication Failure Rates (EFR) which are simply 1 minus the EDR. The decreasing trend for the FDR is significant as the FDR is just a transformation of the EDR. A comparison of the EFR and FDR helps to interpret replication failures in studies with similar sample sizes and power as the original study. The EFR is notably higher than the FDR for both journals and over time, suggesting that replication failures are more likely to be false negatives due to low power rather than false positives in original studies.

Adjusting alpha

Figures 8 and 9 show the impact of lowering the significance criterion, alpha, on the discovery rate. The most notable change occurs when alpha is lowered from .05 to .01. With alpha = .01, about half of all

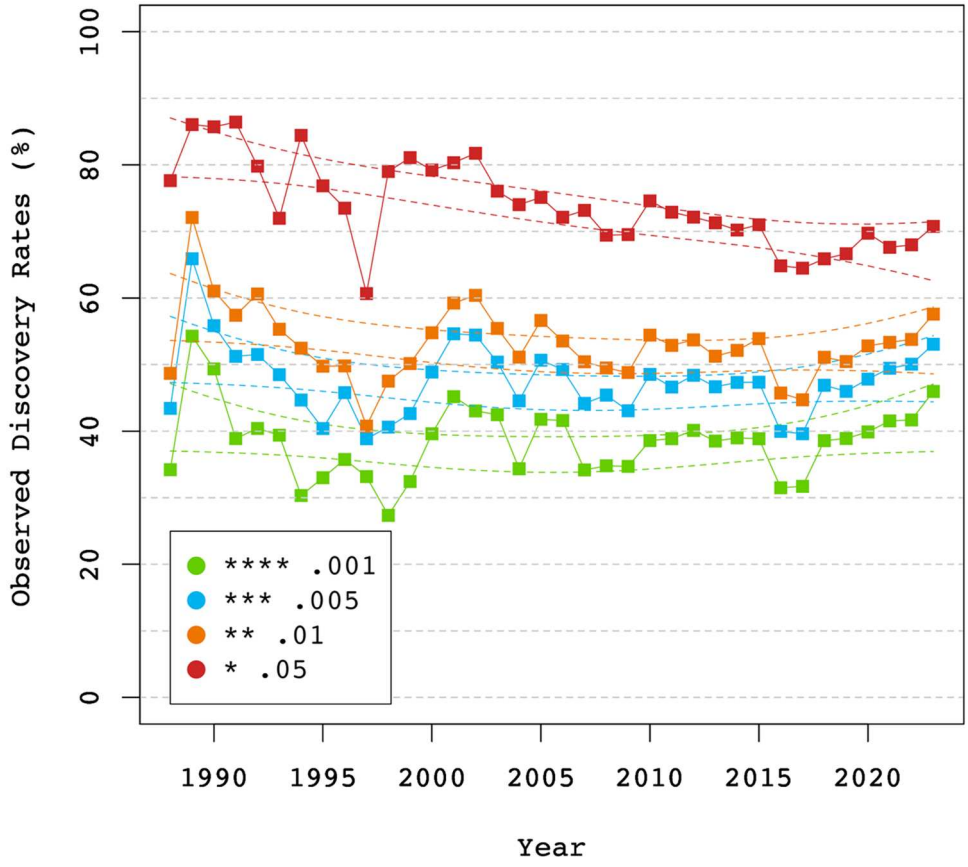

Figure 8. The discovery rate of each alpha level from *Cognition & Emotion*.

published test results remain statistically significant. Figures 10 and 11 show the impact of adjusting alpha on the False Discovery Risk (FDR). Adjusting alpha to .01 is sufficient to reduce the false discovery risk to less than 5% for most years. Further lowering alpha has negligible effects on the false discovery risk. In general, our results suggest that alpha = .01 is the best trade-off between the power to detect true effects and the risk of obtaining false positive results.

Hand coding of focal hypothesis tests

A problem that arises when using automatically extracted data is that not all statistical tests reported are theoretically important. To combat this, we present the results from 241 hand-coded articles published by *Cognition & Emotion* and *Emotion* in 2010 and 2020. The dataset was gathered from an ongoing project with hand-coded focal tests from over 30 journals and over 4000 studies (Schimmack, 2020).

Previous comparisons of automatically extracted results and hand-coded results of focal tests show the biggest discrepancies in the observed discovery rate. As depicted in Figure 12, the ODR was 94% (95% CI [91%; 97%]). Thus, confirming that our ODR estimates of 70% and 71% of significant results underestimate the observed discovery rate for focal hypotheses. In comparison, the EDR, FDR and ERR results remain comparable and well within the confidence intervals of the estimates calculated from the automatically extracted dataset. The main difference arises from the smaller dataset, which leads to greater uncertainty and wider confidence intervals.

Discussion

Emotion researchers are aware that emotions depend on expectations. Our results can elicit different

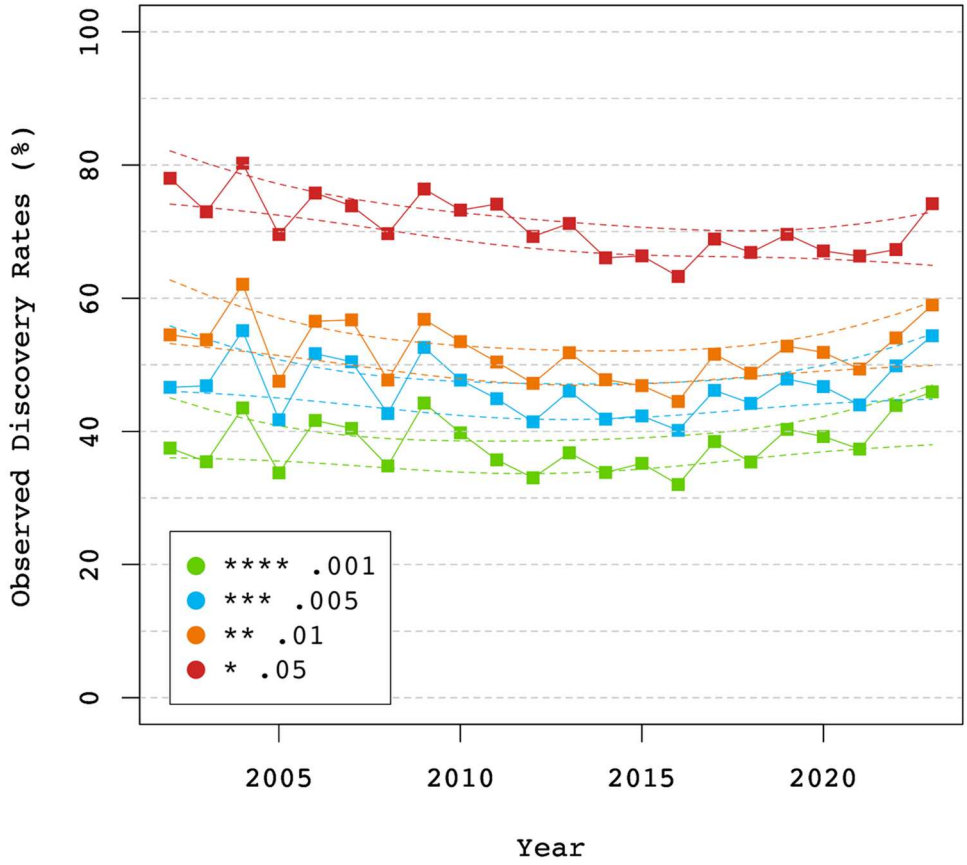

Figure 9. The discovery rate of each alpha level from *Emotion*.

emotions among emotion researchers depending on their prior beliefs about the health of emotional literature. However, after a decade of bad news about the credibility of psychological science, we believe that emotion researchers are likely to feel relief about our findings. In comparison to scenarios that most of the published research is false and that significance is often obtained by employing questionable research practices rather than true effects, our results suggest that only a relatively small percentage of published results are false. Moreover, it is possible to readjust the significance filter to reduce the risk of false discoveries even further. We propose to treat statistical results with *p*-values between .05 and .01 with scepticism. Even multiple replications of a result with *p*-values above .01 do not ensure that the finding is credible. In fact, multiple study articles that have more *p*-values above .01 than below .01 are likely to report results that were obtained with questionable research practices (Schimmack, 2012). In contrast, when *p*-values are consistently below .01, it is unlikely that questionable research practices were used because these practices are more likely to produce *p*-values just below .05 (Simmons et al., 2011). *Z*-curve justifies lowering the general threshold to 0.01 as it caps the false discovery risk below 5%.

Our recommendation is solely intended to be used when reviewing previous literature. We are not proposing a new statistical criterion for the evaluation of new research, although researchers should provide power analyses and justify their alpha level (Lakens et al., 2018). Editors can also be mindful of the fact that *p*-values between .05 and .01 should be rare. If possible, they could ask for additional data to strengthen the empirical evidence for a hypothesis test. Our results also provide only one criterion to evaluate a published article. Readers can combine this information with other information such as preregistration of hypotheses and analysis plans or sample size justifications.

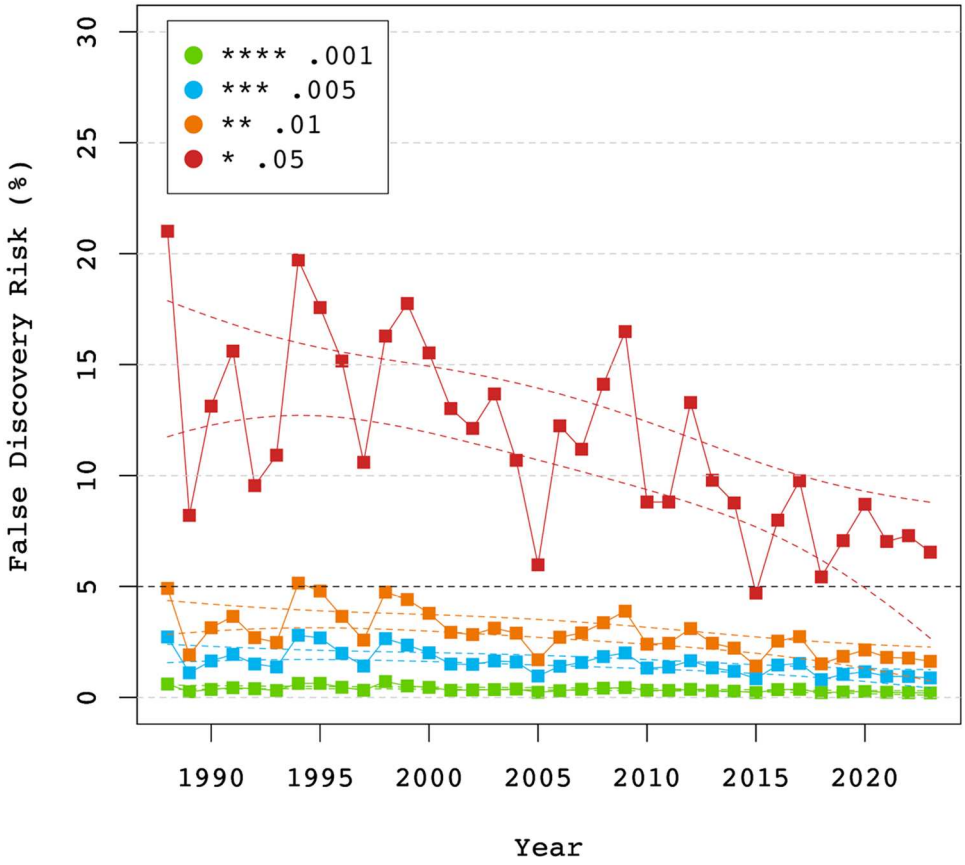

Figure 10. False positive risk of each alpha level from *Cognition & Emotion*.

Our estimate of the replicability of published results is also reassuring. Our results predict a success rate of 65% in replication studies. Moreover, this estimate includes replications of false positives that are assumed to produce a replication failure. Thus, the power to replicate a true finding could be even higher. Assuming an FDR of 10%, power for true hypotheses is 72%, which is close to the recommended power of 80%. However, this estimate is an average. Thus, half of the original studies have less power and require larger sample sizes to avoid false negative results. This estimate tends to be optimistic and the success rate in actual replication studies with the same sample sizes is likely to be lower. Previous work comparing *Z*-curve estimates to replication outcomes suggests that the actual replicability rate tends to fall within the Expected Replication Rate (ERR) and the Expected Discovery Rate (EDR) estimate (Sotola, 2023). Therefore, we expect roughly half of the published results to be replicable. Moreover, replicability is related to the strength of evidence. Results with *p*-values below .01 are likely to produce replication failures unless larger samples are used.

Z-curve can serve as a tool to guide the design of new projects and ensure resources are allocated to replication efforts worth investing in. Based on our analysis, findings with *p*-values above .01 should warrant scepticism. This recommendation may serve as an alternative when running a *Z*-curve analysis is not feasible due to the limited availability of studies on a specific subject of interest. For example, when you observe a pattern of studies consistently reporting *p*-values above .01, it suggests that the evidence is not credible. This was the case for the original studies of the Pen-in-Mouth Paradigm (PIMP), where the pattern expected replication failures (Schimmack & Chen, 2017). If researchers had realised that the

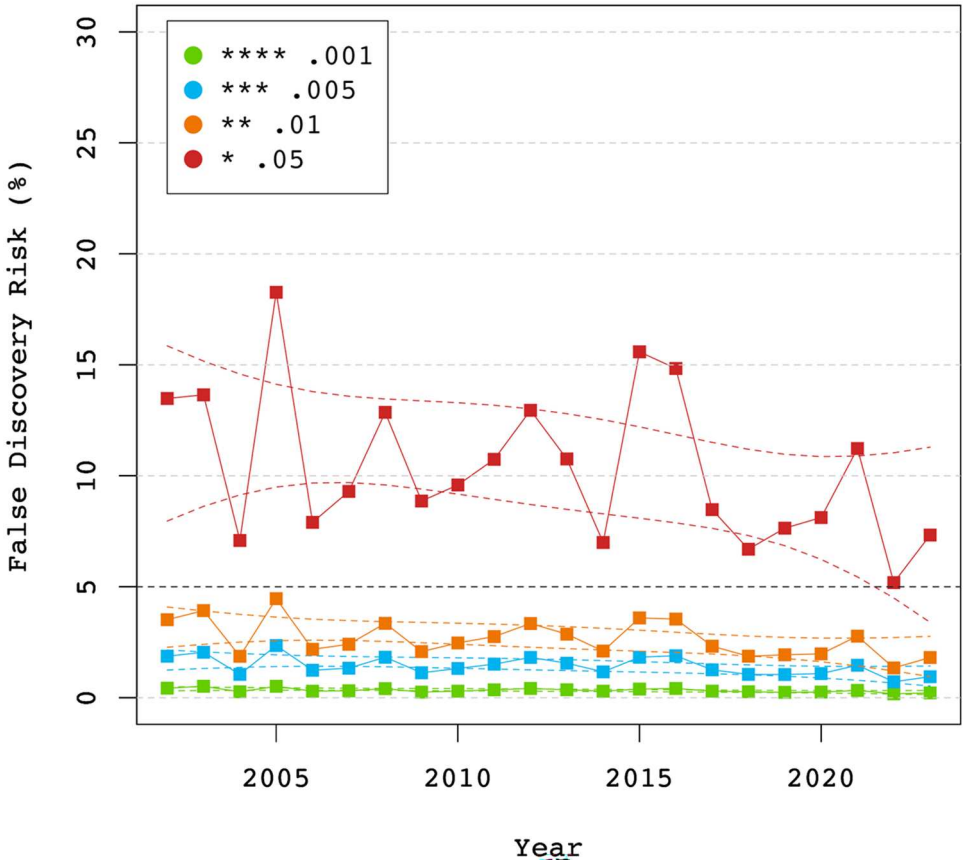

Figure 11. False positive risk of each alpha level from *Emotion*.

Pen-in-Mouth Paradigm (PIMP) never provided convincing evidence, then efforts could have been dedicated toward seeking alternative paradigms to investigate the facial feedback hypothesis. Instead, rigorous registered replication studies were conducted and failed to replicate as expected by the original pattern of *p*-values (Coles et al., 2022; Wagenmakers et al., 2016). Thus, the outcome of the replication studies was not a surprising replication failure, but entirely consistent with the lack of evidence in the originally published studies.

While our results are encouraging for claims about the presence and direction of population effect sizes, the presence of selection bias implies that effect size estimates are often inflated. It is therefore especially important to avoid the interpretation of point estimates of effect sizes and to focus more on the range of plausible effect sizes. When authors do not report confidence intervals, readers should compute their confidence intervals. The interpretation of effect sizes should be avoided when the sampling error is large and selection for significance is required to get statistically significant results. For example, in a between-subject design with $n = 20$ per cell, a standardised effect size of $d = .64$ is needed to obtain $p = .05$ and $d = .85$ is required to obtain $p = .01$. As most effect sizes are smaller than this, the statistically significant results would likely be associated with inflated effect sizes that would be smaller in replication studies. Effect size estimation will often have to rely on meta-analyses of smaller original studies, but meta-analyses of original studies with inflated effect size estimates will also produce inflated effect size estimates. It is therefore important to further decrease publication selection bias by creating a culture that does not impose unrealistic standards of perfection. An article, lab, or journal that only publishes significant results lacks credibility (Schimmack, 2012). In the long run, the discovery rate should match the power of studies to make discoveries.

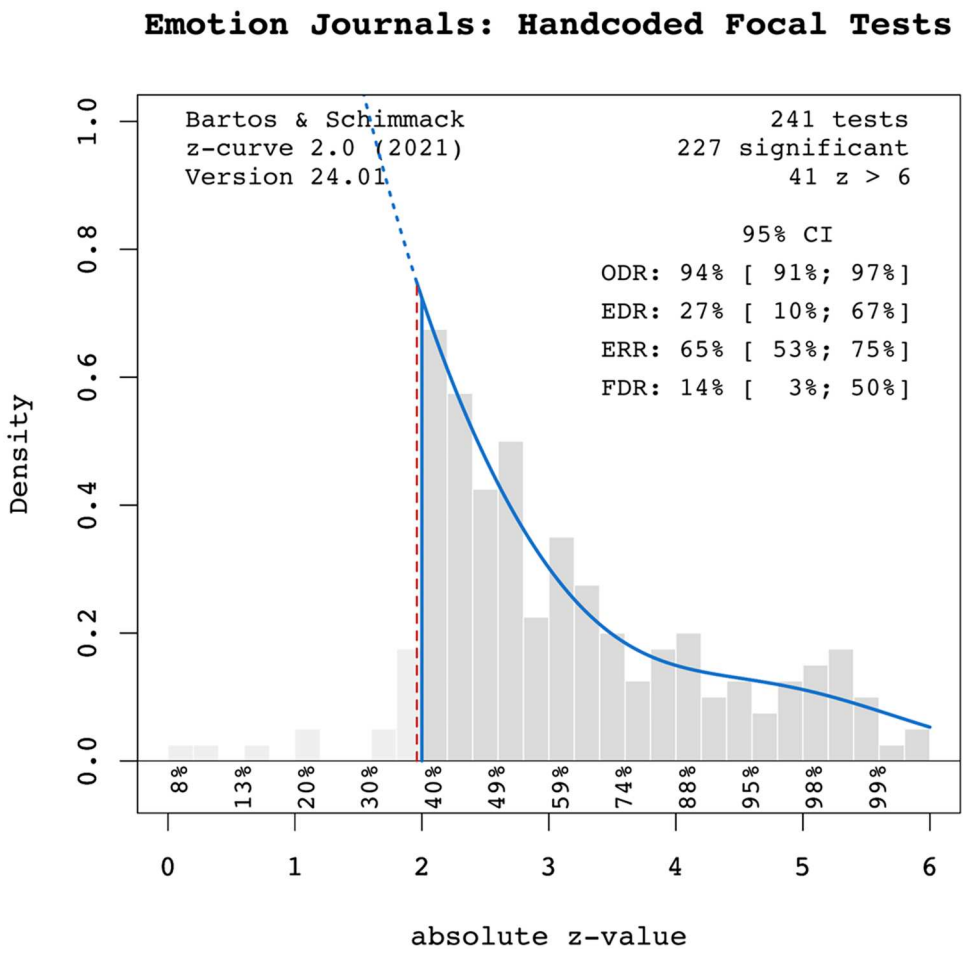

Figure 12. *Z*-curve plot for hand-coded focal tests.

Note: ODR = Observed Discovery Rate, EDR = Expected Discovery Rate, ERR = Expected replication rate, FDR = False Discovery Risk.

Limitations

Like all studies, our study has limitations. *Z*-curve is a selection model that makes assumptions about the selection process. More specifically, *Z*-curve assumes that selection is a simple function of power. However, questionable research practices can produce significance even if true power is low or without a real effect. Simulation studies of *Z*-curve with QRPs are lacking, but some QRPs are likely to produce many *p*-values just below .05. This would produce an inflated estimate of the EDR and an overestimation of selection bias. The problem is that it is unknowable which QRPs are used to obtain better estimates of the EDR. The best solution to this problem is to crack down on the use of QRPs and to report all results. This would produce matching ODRs and EDRs. *Z*-curve analyses of articles published in the coming years can reveal whether emotion researchers are making progress toward this goal.

The present results are limited to articles of designs with over 30 participants, $N > 30$. Further analysis of additional *Z*-curve plots with the complete sample and other exclusions showed similar parameters, this could be because the median degree of freedom for *F*-tests and *t*-tests for the complete sample was 54, ranging from 31 to 104 for *Cognition & Emotion* and 62, ranging from 39 to 106.09 for *Emotion*. However, we recommend that researchers using *Z*-curve to review literature for future research should exclude studies with $N < 30$ or be mindful of possible bias that might be introduced when smaller sample sizes are included. In addition, it would be valuable to run secondary analyses with and without $N < 30$ as a

robustness check. Future work on the method will seek to expand the applicability of *Z*-curve.

Conclusion

Psychology emerged as an empirical science out of philosophy to ensure that theories are grounded in empirical facts. Similarly, meta-science can benefit from empirical evidence. We used the empirical analysis of results published in two emotion journals to examine how credible these results are and what changes emotion researchers might have to make to improve their research practices. Our results show that studies are likely to report the correct sign of a relationship, especially when *p*-values are below .01, but that effect size estimates are inflated. Even meta-analyses will produce inflated effect size estimates because non-significant results are often not reported. We realise that increasing power is not always possible. Therefore, honest reporting of all results is essential to obtain accurate effect size estimates in meta-analyses. To combat selection bias, institutions and journal editors should stop prioritising statistically significant results over non-significant results. Instead, they should reward the relevance of a research question, and the resources used to investigate it. A non-significant result with $N = 200$ can be a bigger scientific contribution than a significant result with $N = 20$.

Acknowledgement

We are grateful for the financial support of this work by the Canadian Social Sciences and Humanities Research Council (SSHRC).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research project was supported by a standard research grant awarded to Ulrich Schimmack by the Canadian Social Sciences and Humanities Research Council. The Open Science Framework repository link is available at https://osf.io/42vxd/.

ORCID

Maria D. Soto http://orcid.org/0000-0001-6825-3985
Ulrich Schimmack http://orcid.org/0000-0001-9456-5536

References

Bartoš, F., & Schimmack, U. (2022). Z curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology* (Växjö), *6*. https://doi.org/10.15626/MP.2021.2720

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology* (Växjö), *4*. https://doi.org/10.15626/MP.2018.874

Cohen, J. (1994). The earth is round (p<.05). *The American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L. G., Willis, M. L., Foroni, F., Reggev, N., Mokady, A., Forscher, P. S., Hunter, J. F., Kaminski, G., Yüvrük, E., Kapucu, A., Nagy, T., Hajdu, N., Tejada, J., Freitag, R. M. K., … Liuzza, M. T. (2022). A multilab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, *6*(12), 1731–1742. https://doi.org/10.1038/s41562-022-01458-9

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), e10068–e10068. https://doi.org/10.1371/journal.pone.0010068

Ioannidis, J P A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, *113*(1), 34–58. https://doi.org/10.1037/pspa0000084

Ooms, J. (2024). pdftools: Text extraction, rendering and converting of pdf documents. [R package version 3.4.0]. https://CRAN.R-project.org/package=pdftools

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian*

Psychology = Psychologie Canadienne, *61*(4), 364–376. https://doi.org/10.1037/cap0000246

Schimmack, U. (2024, November 7). *Questionable reviewer practices: Dishonest simulations*. Replicability-Index. https://replicationindex.com/2024/11/07/questionable-reviewer-practices-dishonest-simulations/

Schimmack, U., & Bartoš, F. (2023). Estimating the false discovery risk of (randomized) clinical trials in medical journals based on published *p*-values. *PLoS ONE*, *18*(8), e0290084–e0290084. https://doi.org/10.1371/journal.pone.0290084

Schimmack, U., & Chen, Y. (2017, September 4). *The Power of the Pen-in-Mouth Paradigm (PIMP): A replicability analysis*. Replicability-Index. https://replicationindex.com/2017/09/04/the-power-of-the-pen-paradigm-a-replicability-analysis/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, *84*(406), 608–610. https://doi.org/10.1080/01621459.1989.10478811

Sotola, L. (2023). How can I study from below, that which is above? : Comparing replicability estimated by Z-curve to real large-scale replication attempts. *Meta-Psychology* (Växjö), *7*. https://doi.org/10.15626/MP.2022.3299

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance-or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108–112. https://doi.org/10.1080/00031305.1995.10476125

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768–777. https://doi.org/10.1037/0022-3514.54.5.768

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458