Faculteit der Sociale Wetenschappen

Replicating the Uncertain

Using Degrees of Freedom Space of Original Articles to Choose Between Studies With a High Replication Value

Celess Datadin

Master's Thesis Methodology and Statistics Master

Methodology and Statistics Unit, Institute of Psychology,

Faculty of Social and Behavioral Sciences, Leiden University

Date: January 19th, 2021

Student number: s1404040

Supervisor: Dr. A.E. van 't Veer

Second reader: Dr. Tom Heyman

Acknowledgments

In front of you lies the thesis ‘Replicating the Uncertain: Using Degrees of Freedom Space of Original Articles to Choose Between Studies With a High Replication Value’. The research regarding the current thesis is conducted in the context of my graduation in ‘Methodology and Statistics in Psychology’ at Leiden University. The research topic is how researchers looking to select a target for replication can use our DFS graphs to map the uncertainty of original work. I chose this topic, because I wanted to contribute to paving the way for replication. Furthermore, I did not know much about replication before commencing this research and I wanted to expand my knowledge through conducting this explorative research. After months of hard work, I can now say that I succeeded in obtaining this goal.

I would especially like to thank Dr. Anna van ’t Veer for her professional supervision during the entire process. Despite the challenges of working from home, I could always call on her expertise. Her experienced judging was evident in the feedback I received, which taught me a lot about both substantive and stylistic aspects of the current thesis. The acquired knowledge will certainly be useful in my endeavors. All in all, I felt more than properly supported under Anna’s guidance. I would also like to thank the second reader Dr. Tom Heyman, and my fellow students Myrthe and Maaike for their input. Finally, I would like to express my gratitude to my boyfriend Pim, my mother Indra, my sister Dainara, and my brother-in-law Charly: thank you all for your unconditional love and support, and for providing me with a comfortable home office.

I hope you enjoy reading the final product!

Celess Datadin

Leiden, 19 January 2021

Abstract

Flexibility in the decisions researchers make during their research can lead to false positive findings. Due to low transparency in published papers in the field of psychology, the amount of flexibility authors had is often unclear. In the current thesis, in a first step a quantitative measure of Replication Value is applied to a random set of studies (n = 1257) from Social Psychology, using citation count as a proxy for impact and sample size as a proxy for uncertainty. This Replication Value has been suggested as an indicator of how worthwhile it is to replicate a study (see Isager et al., in press), and can be applied to a large number of studies due to its quantitative approach. However, Replication Value is based on solely on quantitative proxies. Therefore, it is necessary to also manually examine papers. In a second step of the current thesis, it is manually explored whether the uncertainty that researchers have when making choices during their research can become clearer by mapping them. Therefore, the studies with the highest Replication Values (n = 10), with the median Replication Values (n = 10), and with the lowest Replication Values (n = 10) were examined on their reporting transparency and potential Researcher Degrees of Freedom. A detailed analysis of the first results indicated that the qualitative analysis of the Researcher Degrees of Freedom of original researchers is helpful to in selecting which study to replicate after making a larger selection based on RV. The findings from this exploratory research are discussed in the context of the field of Social Psychology, with an emphasis on how researchers looking to select a target for replication can use our DFS to map the uncertainty of original work.

Keywords: replication, Questionable Research Practices, Researcher Degrees of Freedom, social psychology, transparency

Table of Contents

Replicating the Uncertain

Science is often associated with discovery: finding a new and exciting phenomenon. This excitement arguably comes at a high cost, namely the neglect of the self-correcting element of science where past findings are examined for robustness and existing knowledge is continually updated. In the current academic world, there is an emphasis on new and original studies and findings: Researchers feel the need to make new discoveries and journals are not likely to publish replication studies. Thus, neither is properly incentivized to invest time, money and/or energy in replicating previously conducted studies or making their own studies accessible for replication. The current thesis examines the selection process of which studies are in need of replication. This selection process is broken down in two parts, where a quantitative formula that can be applied to a large number of candidates is combined with a more detailed examination of candidates suggested by this formula. This latter examination aims to approximate the uncertainty surrounding the original researcher's decision flexibility. If proven informative, this measure of uncertainty can be utilized by researchers looking to select a candidate for replication.

There are several reasons why replicating an existing study is actively discouraged in the field of (social) psychology. On the one hand, researchers feel the need to make new discoveries (Makel, Plucker, & Hegarty, 2012). They are rewarded for reporting novel findings in the form of a more prestigious and better reputation within the academic community (Ebersole et al., 2016), a better chance of getting the paper published (Nosek, Spies, & Motyl, 2012), increased odds of being cited and getting favorable peer reviews (Joober, Schmitz, Annable, & Boksa, 2012), and more funding for future research since replication studies are not likely to get funded by funding agencies (Artino Jr., 2013). On the other hand, journals are not likely to publish replication studies (Makel, Plucker, & Hegarty, 2012). Journal editors and reviewers are inclined to disfavor replication studies (Spellman, 2012), because novel, statistically significant results are attention-grabbing (Ebersole, Axt, & Nosek, 2016) and thus likely to generate more subscriptions and citations (Joober et al., 2012). More citations lead to a higher ranking, which attracts more paying subscribers. Generating revenue is one of the reasons for journals to discourage replication studies (Buranyi, 2017). Furthermore, some journals even have a policy against replications (Ritchie, Wiseman, & French, 2012). These aspects of the scientific (incentive) system together create a problematic lack of replication in the field of Social Psychology.

The literature is thus full of novel findings, yet scientific progress does not rely solely on novel findings. This notion is expressed very aptly in the following quote: "Innovation points

out paths that are possible; replication points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not" (Open Science Collaboration, 2015, p. 943). Currently, journals do not deem replications valuable to the progress of science in a specific field (Block & Kuckertz, 2018). Moreover, two thirds of the 1576 surveyed researchers from a *Nature* survey do not think that failed replications indicate wrong published results (Baker, 2016). Nevertheless, because of a lack of replication studies, it is unknown how reliable findings in a field are. The disproportional emphasis on positive over negative results causes an inflated false positive rate in published papers (Nosek et al., 2012; Chambers, 2017). Therefore, more replications are needed to test scientific findings for robustness and to better estimate the certainty with which they can be relied on.

One reason for the doubts about the reliability of the field of Social Psychology is the presence of researcher degrees of freedom (RDF; Simmons, Nelson, & Simonsohn, 2011) in many original findings. RDF entail the large number of decisions made by researchers during data collection and analysis. Because constraining this freedom via preregistration is relatively new, studies in the existing literature may have uncertainty about the amount of flexibility of an original author; uncertainty that arguably can lead to a higher need to replicate said study in order to better estimate the reported effect. In this thesis, this uncertainty surrounding the original authors' room for flexibility is called 'degrees of freedom space', and this space will be examined to see whether it aids the selection process of a target for replication beyond the aforementioned quantitative approach.

There are many reasons why having the flexibility to make ad hoc decisions can cause uncertainty about reported results. Because of the room for flexibility, it is possible for researchers to engage in different Questionable Research Practices (QRPs; John, Loewenstein, & Prelec, 2012). QRPs fall between responsible conduct of research (RCR; Steneck, 2006) and fabrication, falsification, and plagiarism (FFP; Steneck, 2006). The three most prevalent QRPs among academic psychologists are estimated to be: failing to report all dependent measures, collecting more data after seeing whether results were significant, and selectively reporting studies that 'worked' (John et al., 2012; see Appendix A for an overview of QRPs and RDF). The room for flexibility is described by Gelman and Loken (2014) as a garden of forking paths in which implicit choices are made by researchers. Simmons and their colleagues (2011) showed through simulations and experiments that room for flexibility can lead to a dramatical increase of false positive findings. These false positives make a successful replication unlikely, and the decision flexibility that original authors had and their often untransparent way of

reporting about the chosen route make it unlikely a replicator will be able to decipher how original results were obtained.

Because replication is one of the possible ingredients of the much-needed paradigm shift in the field of (social) psychology, deciding which replication studies are worth our resources (e.g., time and money) is an important matter. One of the main reasons for this is that it is not possible to replicate every single study, but nonetheless it is desirable to have more certainty about the reliability of this field. Using a quantitative formula to calculate a Replication Value (RV) which aids researchers in choosing which findings to replicate, is fruitful because it is neither possible nor efficient to replicate all findings. In order to determine which findings are more worthwhile to replicate, Isager and colleagues (in press) created a formula – see (1) – that takes into account two relevant characteristics of findings: importance and certainty.

On the one hand, the findings should be important, because important findings are assumed to have more impact and consequences than less important findings. In the current thesis, citation count is used as a proxy for the difficult to measure concept of importance. The reasoning behind this is that citation count takes into account some sort of academic consensus about the importance of findings (Bastow, Dunleavy, & Tinkler, 2014). The downside is that citation count does not take into account the wider influence of findings in external communities outside the academic one, such as public policy, media, cultural, civil society, economic, and business systems (Bastow et al., 2014). Despite these downsides, the RV formula still uses citation count as a proxy for impact, because it is a straightforward metric that is relatively easy to obtain for large amounts of papers at a time.

On the other hand, uncertain findings will most likely lead to the most essential replications, whereas certain findings are in less need of more evidence. Isager (2019) has operationalized certainty, or 'corroboration', as estimation precision, which is quantified as the variance of Fisher's *Z*. The variance of Fisher's *Z* is only dependent on sample size (Isager, 2019). In the current thesis, therefore, the extent to which a finding is uncertain is also measured by sample size. The assumption is that a finding that is based on a small sample size is more uncertain than a finding that is based on a larger sample size. A small sample is less representative of the entire population, and less able to detect statistically significant differences (Verma & Verma, 2020) while more likely to lead to a false positive (Button et al., 2013).

Equation (1) shows how to calculate the RV by dividing the total citation score (*TC*) by the sample size after exclusion (*SS*), after correcting for the years since publication (*PY*).

$$RV = \frac{Impact}{Corroboration} = \frac{TC}{PY + 1} * \frac{1}{SS} \quad (1)$$

First, the total citation score is in the nominator of (1), because more citations are assumed to indicate that the finding has a larger impact. Therefore, the replication value is higher when the citation score is higher. Next, the citation score is divided by how many years have passed since the publication year. The objective of this division is to take into account that newer papers (i.e., papers with a lower number of years since publication) will have had less time to get cited as opposed to older ones (i.e., papers with a higher number of years since publication). The last step of (1) is to divide the result of the former fraction by sample size (i.e., multiplying with 1 divided by sample size), because RV is inversely proportional to the sample size. A lower sample size indicates a lower amount of certainty about the finding. Taken together, this leads to a highly cited study from which an uncertain finding stems, to be in higher need of replication.

Equation (1) is a quantitative approach for selecting studies which can be applied to a large number of candidates. After ranking this large number of candidates, a qualitative approach can be applied where researchers looking to select a target for replication can manually go through the top ranked candidates. A possible aid in this process is a measure of uncertainty surrounding the RDF of the original work. In the current thesis, therefore, the 'degrees of freedom space' of original papers is evaluated in its ability to aid the selection process of what to replicate. This space is mapped by investigating the extent to which original papers leave ambiguities concerning known 'grey' areas, wherein QRPs can possibly take place.

Next to examining whether the 'degrees of freedom space' of the original paper can aid the selection of one replication study from the studies with a high RV rank, it is expected that the ranking based on quantitative indicators (i.e., sample size and citation score) results in a top of studies that are deemed more worthy of replication than the center or bottom ranked studies after analyzing them with a focus on the 'degrees of freedom space' of the original researcher(s). In other words, if mapping the 'degrees of freedom space' for the top 10 studies results in a bigger space compared to the center or bottom 10), this would to add to the utility of using this approach to complement the initial RV ranking.

The current thesis aims to address two research questions:

- What are the characteristics (both similarities and differences) of the top 10 of a ranking of studies based on a Replication Value (which is in turn based on sample size and

citation score) compared with center and bottom ranked studies when looking at the 'degrees of freedom space' of the original work?

- The second question is: Can certain characteristics be used to map the original researcher's 'degrees of freedom space' – an indicator of the reproducibility of the decisions made by the researcher(s) – in order to aid in the selection of the paper that is most worthy of replication?

As comparing characteristics of papers in the entire literature base of a field is not humanly possible, this research commences with describing the process of taking a random sample out of the field of Social Psychology. Then, the representativeness of this sample is evaluated based on bibliometric indicators, the quantitative ranking of what is worthy to replicate is applied, and finally the use of the 'degrees of freedom space' of the top ranked studies is evaluated to manually select a target for replication. To ensure this selection process is similar to a situation where a researcher is looking to replicate a study, the aim is to actually select one study and, given this can be accomplished through the described process, use this study as a replication target in further research.

Methodology

The current research is exploratory, because its aim is to assess the utility of combining RV with 'degrees of freedom space' (hereafter: DFS) for making decisions about what to replicate. A mixed methods design has been applied by combining a quantitative and qualitative research component (Creswell & Clark, 2007). The quantitative component consists of a formula – see (1) – by Isager and colleagues (in press) to calculate RV which aids researchers in choosing which findings to replicate. The qualitative component of the method is to determine the utility of DFS as a measure of uncertainty surrounding the original researcher's decision flexibility that can be assessed by researchers looking to select a candidate for replication.

All analyses are executed with *R* (version 4.0.3, R Core Team, 2013). *R* code is provided in Appendix E. The current thesis commences with creating a dataset existing of randomly sampled papers (n = 999) from a total pool of 150.000 papers within the field of Social Psychology. For 961 studies of that random sample, the sample sizes of all their reported studies have been previously coded. Furthermore, 57 papers initially had an unknown DOI, and these were manually added. The RV was calculated for all observations in the dataset (n = 1257) based on their total citation score and sample size after exclusion. Thereafter, all studies from the dataset were ranked based on their RV. Then, the top, center, and bottom studies (n = 30)

with respectively the highest, median, and lowest RVs are selected from the data. Next, on the basis of known researcher decision flexibilities and the extent to which these are traceable in original papers, a list of DFS items was created. Finally, the DFS of the top, center and bottom original studies is evaluated in its ability to aid the selection process of what to replicate. Aforementioned qualitative evaluation of the DFS is done by investigating the extent to which differences exists between top, center and bottom original studies in terms of their ambiguities concerning known 'grey' areas, and whether this subjectively helps pinpoint the most uncertain finding for replication.

Operationalization of RV Ranking

In order to determine which original findings are worthwhile to replicate, it is essential to somehow quantify the expected utility of potential replications (Coles, Tiokhin, Scheel, Isager, & Lakens, 2018). Equation (1) is a way to calculate a single number that encompasses such an expected utility, namely RV. RV indicates to what extent a study is worthwhile to replicate by taking into account two relevant characteristics of findings, namely their importance and certainty (Isager et al., in press). It is important to keep in mind that the formula for RV is an approximation, because the importance and certainty of studies is measured by proxy's.

The importance of findings is operationalized by using the total citation score as a proxy. The Times Cited Count (TC) field tag of Web of Science (WoS; "Web of Science Core Collection Help," 2020) is used as citation score variable, because the current sample of papers is extracted from WoS. Although citation scores do not take into account the wider influence of findings in external communities outside the academic one, they are an indicator of some sort of academic consensus about the importance of findings (Bastow et al., 2014). As Waltman and Noyons (2018, p. 4) state: "They do not provide exact measurements of scientific impact, but they do offer approximate information about the scientific impact of publications, researchers, or research institutions." Equation (1) implicitly assumes that the total citation score of a paper positively correlates with its value of replication (i.e., a higher citation score increases the value and vice versa), because important findings are argued to be more worthwhile to replicate than less important findings – even though scientific impact does not necessarily equal practical impact (i.e., real world consequences). Equation (1) shows that the total citation score is divided by the number of years that have passed since the paper was published. This acts to correct for relatively new papers having had less time to get cited as opposed to older papers.

The certainty of findings is operationalized by using the total sample size after exclusion as a proxy, because estimation precision is quantified as the variance of Fisher's *Z* which is only dependent on sample size (Isager, 2019). RV is inversely proportional to the sample size of a paper (i.e., a larger sample decreases RV and vice versa), because uncertain findings are argued to be more worthwhile to replicate than more certain findings. The sample size after exclusion has been previously coded for each study within a paper.

Sample and Procedure

Description of initial sample. The master file (i.e., the sample of 999 distinct papers) came about by taking a random sample from a pool of ~ 150.000 papers within the field of Social Psychology. This sample is taken from the WoS database. The merged dataset (*n* = 1257) results from the merge of the master file and the extra dataset containing 256 extra studies from the extra dataset for those of the 999 papers that reported more than one study. The sample sizes of all studies are coded by five different coders, where 10% was double (or in some cases) triple checked to ensure reliable coding. The data were cleaned by completing 35 missing DOI's and 15 missing sample sizes after manually looking them up (see Appendix E for full *R* code containing all cleaning steps). Not every paper in the data reports the same number of studies: three papers report six studies, three papers report five studies, seven papers report four studies, 35 papers report three studies, 88 papers report two studies, and 741 papers report one study.

Inclusion criteria for current purposes. As shown in Figure 1, six exclusions have been made from the merged dataset (*n* = 1257) to create a final dataset. Firstly, papers with an unknown DOI (*n* = 21 in the merged dataset) are excluded from further analyses, because those papers also miss information on a lot of other relevant variables (e.g., citation score and sample size). Secondly, papers that are published later than 2018 (*n* = 34 in the merged dataset) were excluded, because they arguably did not get enough time to be cited. Thirdly, RV cannot be calculated for papers with an unknown citation score (*n* = 42 in the merged dataset), unknown sample size (*n* = 21 in the merged dataset), and/or unknown publication year (*n* = 0 in the merged dataset). Therefore these papers are also excluded. Finally, for each paper, only the study with the largest sample size is selected, because it is assumed that in the field of social psychology selecting e.g. the first study would bias the ranking as these first studies are often pilot studies and therefore have a lower sample size. If a paper reports about one single study, then this study automatically is coded as the study with the largest sample size of said paper. This final condition leads to excluding studies that do not have the largest sample size within a paper (*n* = 244 in the merged dataset). Thus, the final dataset (*n* = 937), for which the RV was

calculated, does not contain any papers that are published in or after 2018 and/or have missing citation scores, sample sizes, and/or publication years. Furthermore, the final dataset contains only those studies that have the largest sample size within each paper. After successfully calculating the RV for all studies in the final dataset (n = 937), the studies were ordered from highest to lowest RV.

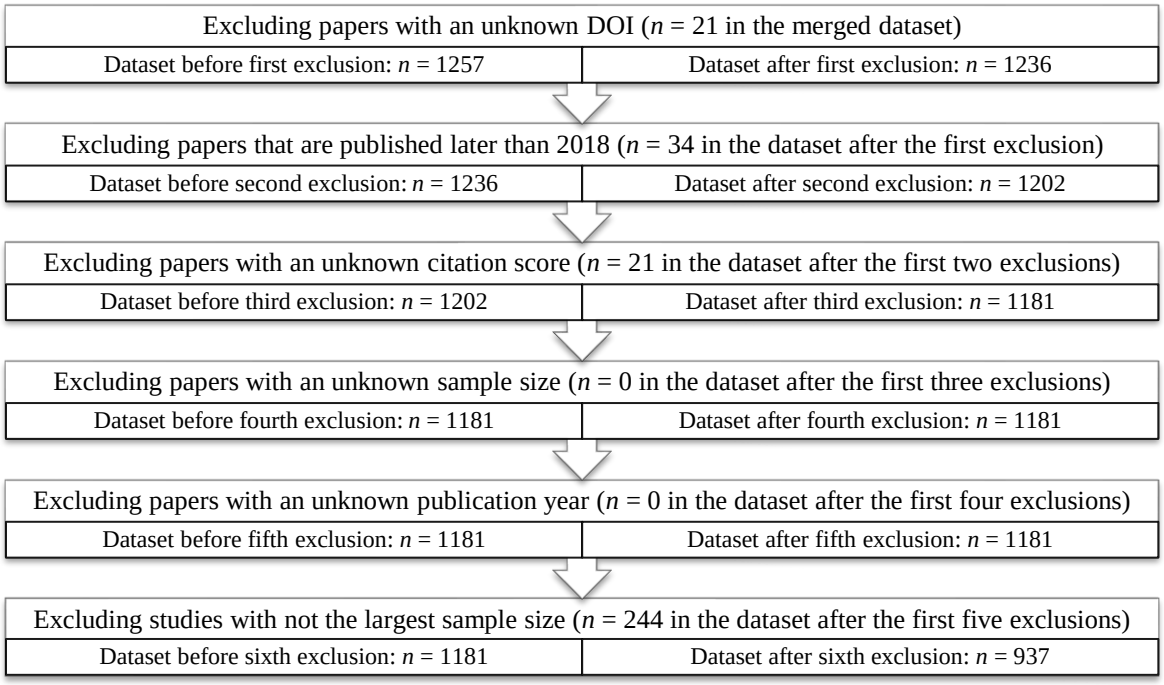


Figure 1. Exclusions to create the final dataset (n = 937).

Journal descriptives. In what follows, first characteristics of the journals of all studies (i.e., before the sixth exclusion in Figure 1) are described, followed by a description of the characteristics of the journals of the study, or in case of multiple studies per paper, the study with the largest sample size within the final dataset (i.e., after the sixth exclusion in Figure 1). For all studies within each paper, the frequencies of the journals with at least ten articles (n = 768) are shown in Figure 2a. The three most prevalent journals in the data with all studies are *Personality and Individual Differences* (n = 99), *Journal of Personality and Social Psychology* (n = 87), and *The Journal of Social Psychology* (n = 69).

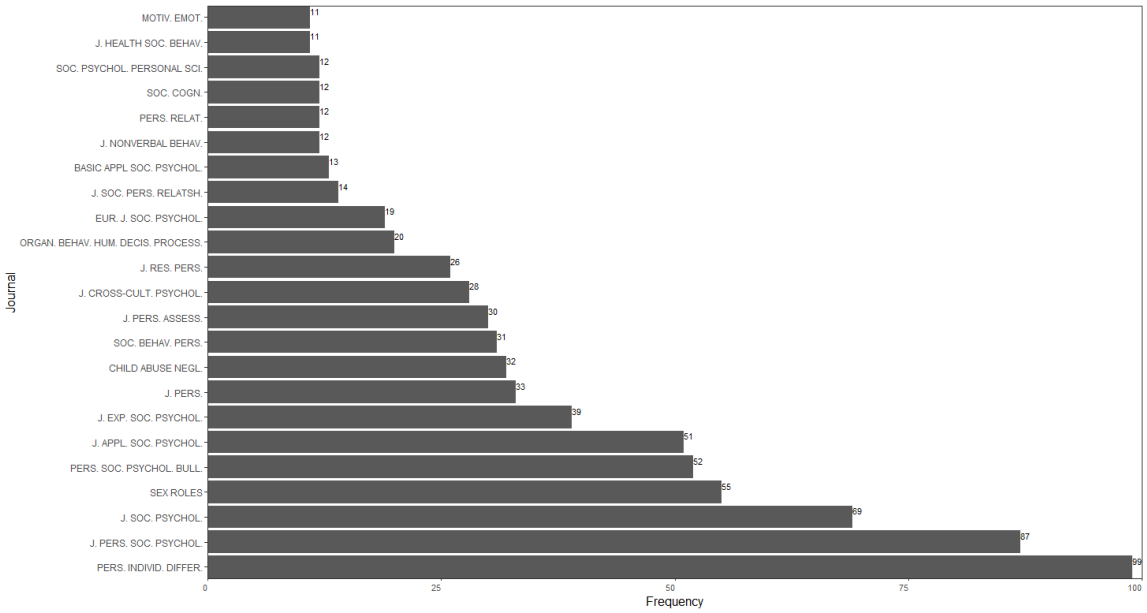


Figure 2a. Frequency of journals with at least ten articles (*n* = 768) in the data with all studies (*n* = 1181).

For only the studies with the largest sample size within a paper, the frequencies of the journals with at least ten articles (*n* = 685) are shown in Figure 2b. The three most prevalent journals in the data with only the largest studies per paper are the same as in the data with all studies: *Personality and Individual Differences* (*n* = 96), *Journal of Personality and Social Psychology* (*n* = 70), and *The Journal of Social Psychology* (*n* = 69).

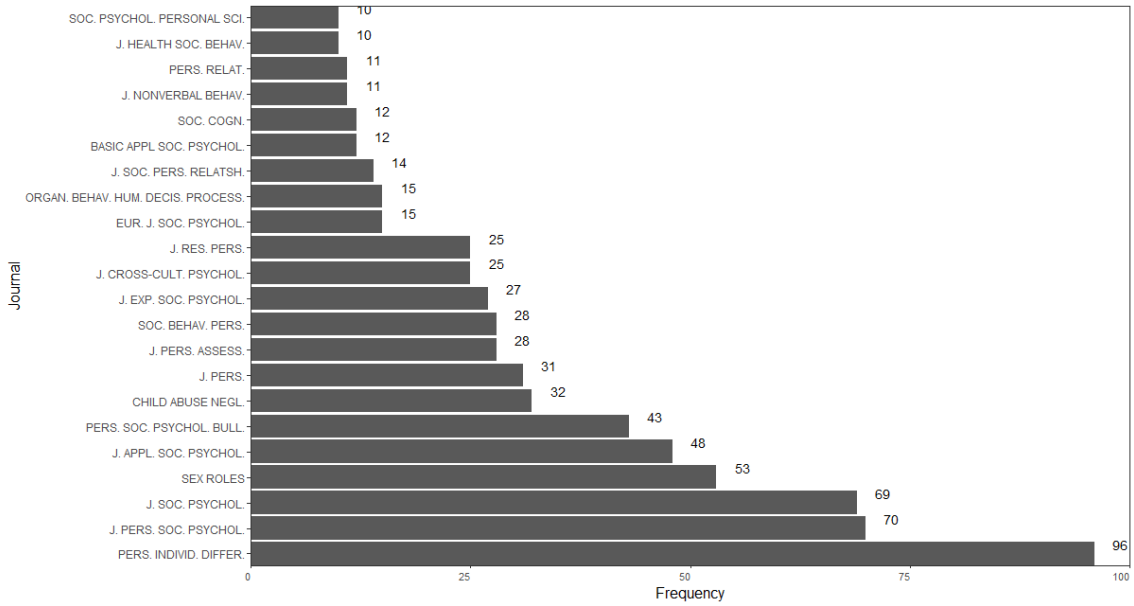


Figure 2b. Frequency of journals with at least ten articles (*n* = 685) in the data with only the largest studies (*n* = 937).

Table 1 (taken from Sassenberg & Ditrich, 2019) shows the "Mean Sample Size, Mean Percentages of Studies Using Online Data Collection and Only Self-Report Measures, and Mean Number of Studies per Article, by Journal and Publication Year" (p. 111). Sassenberg and Ditrich (2019) chose the four journals shown in Table 1, because they are "the four top empirical social psychology journals" (p. 108). The *Journal of Personality and Social Psychology* the second top social psychology journal (Sassenberg & Ditrich, 2019; see Table 1), and is the second most prevalent in the both the data with all studies (n = 87) and the data with only the largest studies (n = 70). *Social Psychology and Personality Science* is the fourth top social psychology journal (Sassenberg & Ditrich, 2019), but is one of the least prevalent journals in both the data with all studies (n = 12) and the data with only the largest studies (n = 10). The first and third top empirical social psychology journals (Sassenberg & Ditrich, 2019) are also part of both the data with all studies and the data with only the largest studies: *Journal of Experimental Social Psychology* (n = 39 and n = 27) and *Personality and Social Psychology Bulletin* (n = 52 and n = 43). Thus, except for *Social Psychology and Personality Science* (n = 12 and n = 10), the frequencies of social psychology journals in the sample (both before and after selecting only the studies with the largest sample size within each paper) seem to be representative of the field.

Table 1

Copy of Table 3 from Sassenberg & Ditrich (2019, p. 111)

Variable and year	*JESP*	*JPSP*	*PSPB*	*SPPS*
Sample size				
2009	113 (98)		122 (115)	
2011	112 (98)	102 (72)	138 (99)	130 (111)
2016	142 (106)	195 (120)	180 (131)	198 (158)
2018	203 (134)		185 (115)	
Overall	145 (117)	145 (108)	161 (120)	165 (141)
Online data collection (%)				
2009	9.0		2.6	
2011	11.6	5.0	12.2	18.7
2016	33.7	64.2	37.7	42.3
2018	49.2		50.5	
Overall	26.6	32.6	28.2	30.9
Only self-report measures (%)				
2009	43.8		48.7	
2011	33.1	44.6	40.9	36.3
2016	32.7	76.4	57.9	66.0
2018	71.7		63.8	
Overall	46.0	59.5	53.4	51.6
Studies per article				
2009	1.96 (1.17)		2.62 (1.21)	
2011	2.55 (1.21)	3.78 (1.60)	2.43 (1.03)	1.88 (0.93)
2016	3.09 (1.25)	5.94 (1.84)	3.25 (1.64)	2.13 (1.35)
2018	3.24 (1.40)		3.79 (1.62)	
Overall	2.65 (1.34)	4.50 (1.96)	2.98 (1.47)	2.00 (1.15)

Sample size descriptives. In what follows, characteristics of the sample sizes of the study, or in case of multiple studies per paper, the study with the largest sample size within the final dataset are described. The distribution of the sample sizes (*M* = 396) is shown in Figure 3a. The largest study has a sample size of 29472 and the smallest study has a sample size of 8.

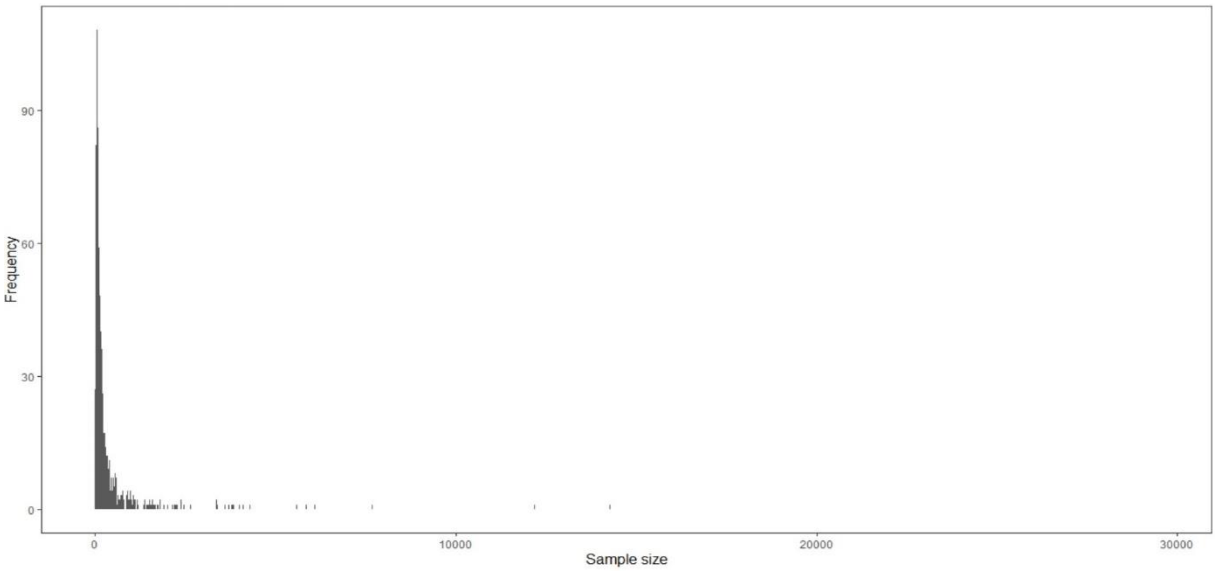


Figure 3a. Distribution of all sample sizes in the final dataset (*n* = 937).

In order to give a more detailed picture of the distribution of sample sizes, Figure 3b shows how frequent the sample sizes of 500 and lower are the data with only the largest studies. The most prevalent sample size was 40 (*n* = 21), followed by 60 (*n* = 19) and 80 (*n* = 17).

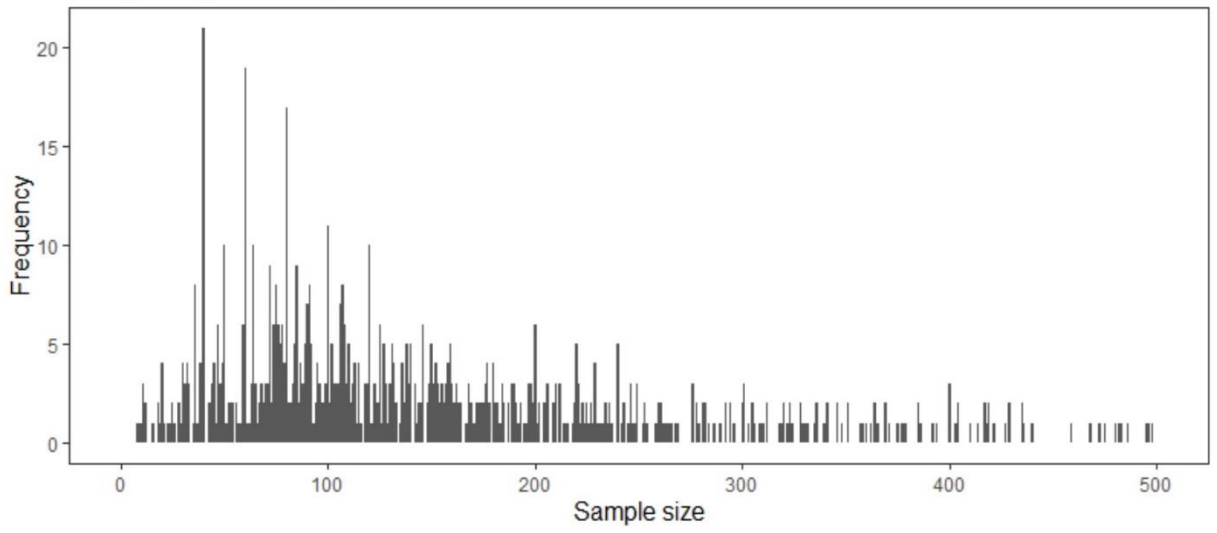


Figure 3b. Distribution of sample sizes smaller than or equal to 500 (*n* = 798) in the final dataset (*n* = 937).

As shown in Figure 4, the largest sample size in the most papers belongs to the first study that is reported (n = 847), followed by the second (n = 60) and third reported study (n = 22).

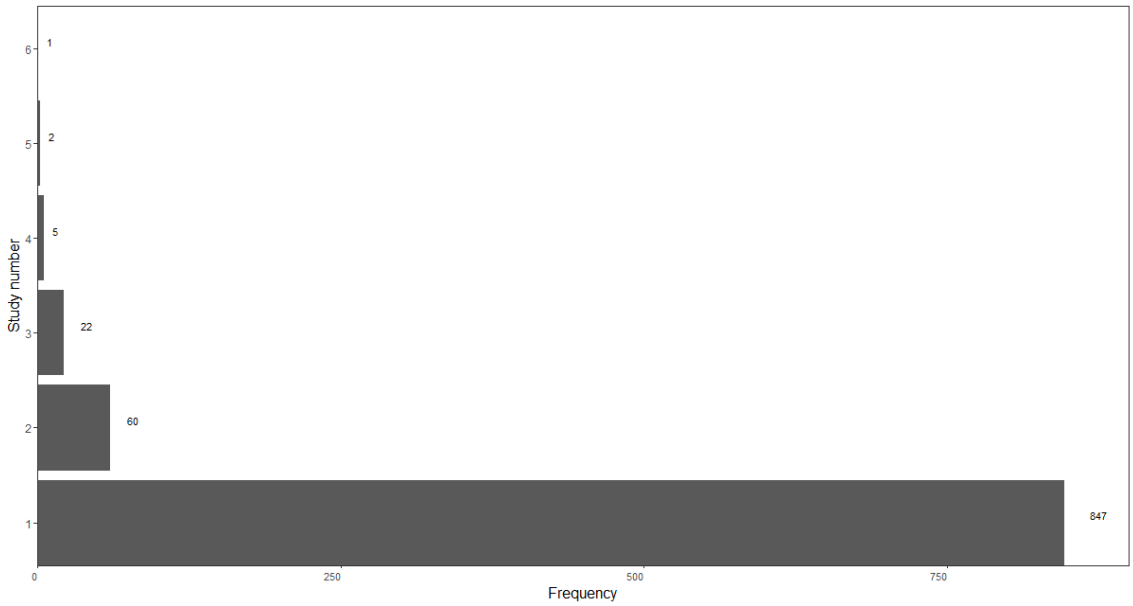


Figure 4. Frequency of all study numbers in the final dataset (n = 937).

Citation count descriptives. In what follows, characteristics of the citation count of the study, or in case of multiple studies per paper, the study with the largest sample size within the final dataset are described. The distribution of the citation scores (M = 32) is shown in Figure 5a. The most cited paper has a citation score of 842 and 50 papers have 0 citations.

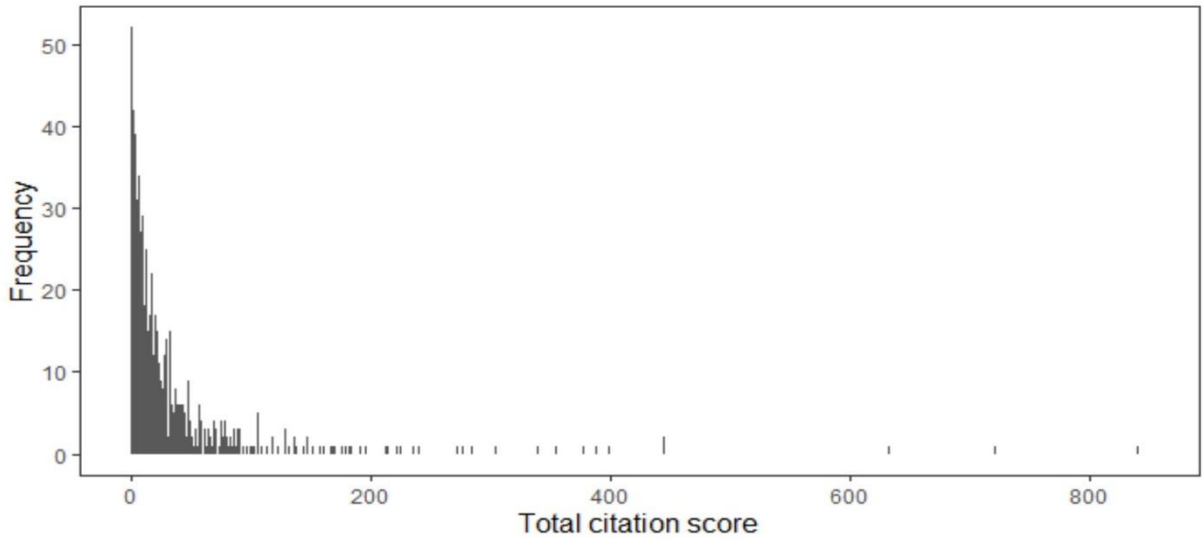


Figure 5a. Distribution of all citation scores in the final dataset (n = 937).

In order to give a more detailed picture of the distribution of citation scores, Figure 5b shows the frequency of the citation scores that appear at least 2 times in the data with only the largest

studies. The lowest citation scores are displayed on the left and the highest on the right. The highest bar shows that 52 papers are cited one time.

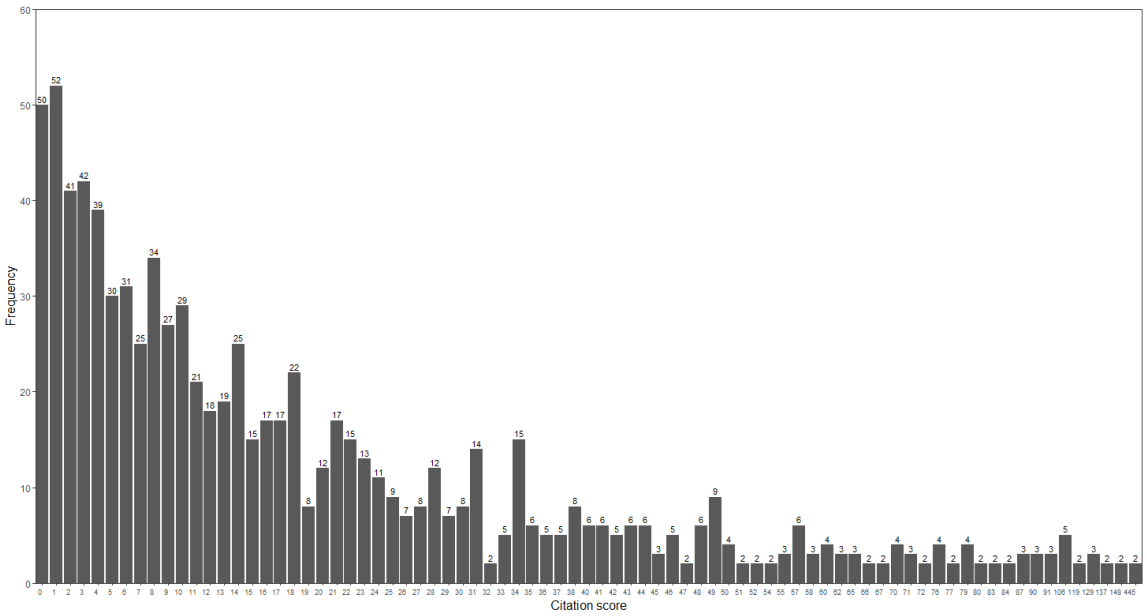


Figure 5b. Frequency of citation scores with at least a frequency of 2 (*n* = 881) in the final dataset (*n* = 937).

In order to gain insight into the extent to which citation score is a suitable indication for RV, the correlations between sample size, years since publication, citation score, and RV are shown in Figure 6a. The Pearson's correlation coefficients between all variables that are used in equation (1) are small (< |.2|), except for the correlation between citation score and RV (*r* = .57).

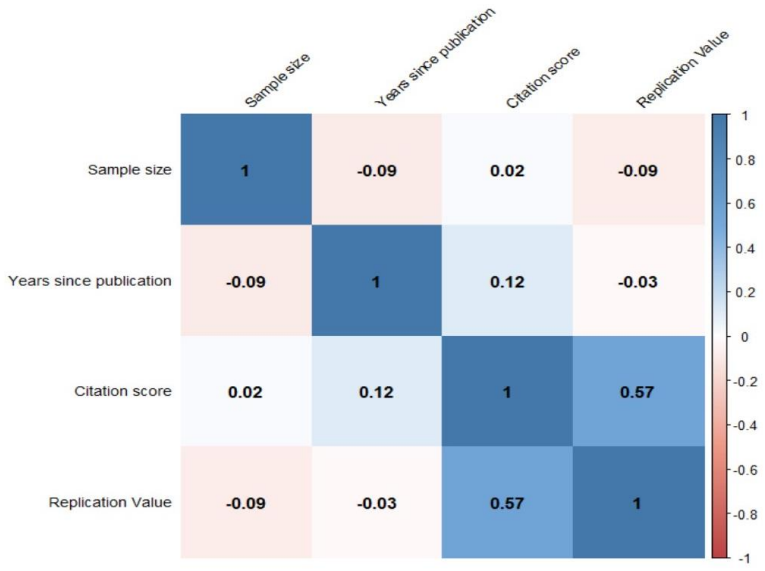


Figure 6a. Correlations between four relevant variables in the final dataset (*n* = 937).

As shown in Figure 6b (left), the linear regression line suggests a positive linear relationship between RV and citation score (r = .57). The top 10 studies have especially high citations scores, as shown by the blue line in Figure 6b (right).

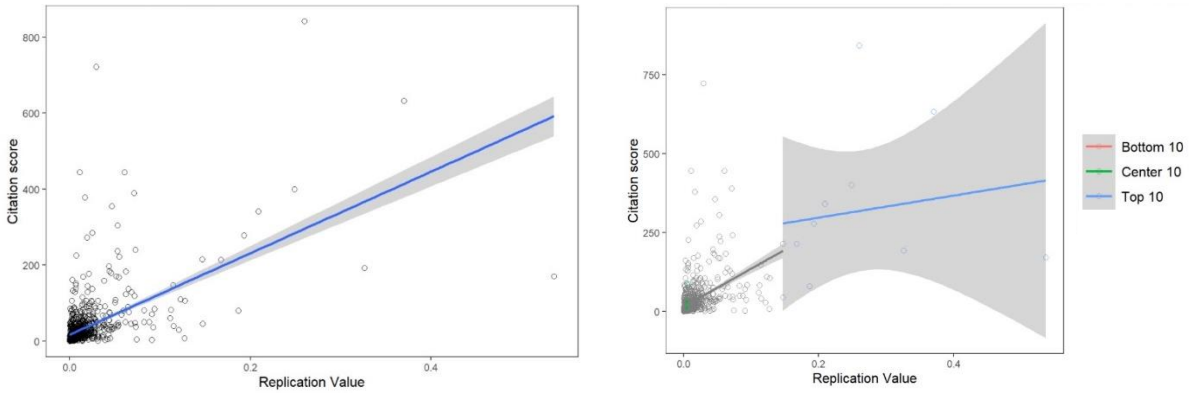


Figure 6b. Scatterplots of RV and citation score in the final dataset (n = 937).

Publication year descriptives. In what follows, characteristics of the publication year of the study, or in case of multiple studies per paper, the study with the largest sample size within the final dataset are described. The distribution of the years of publication (M = 1999) is shown in Figure 7. As can been seen, it is left skewed, with more recently published papers than early published ones. In the data with only the largest studies, the oldest paper was published in 1949 and the newest in 2018. Most articles were published in 2010 (n = 41), followed by 2018 (n = 39) and 2015 (n = 38). The growing amount of published papers is aligned with the overall trend in the field of social psychology (Cutting, 2007).

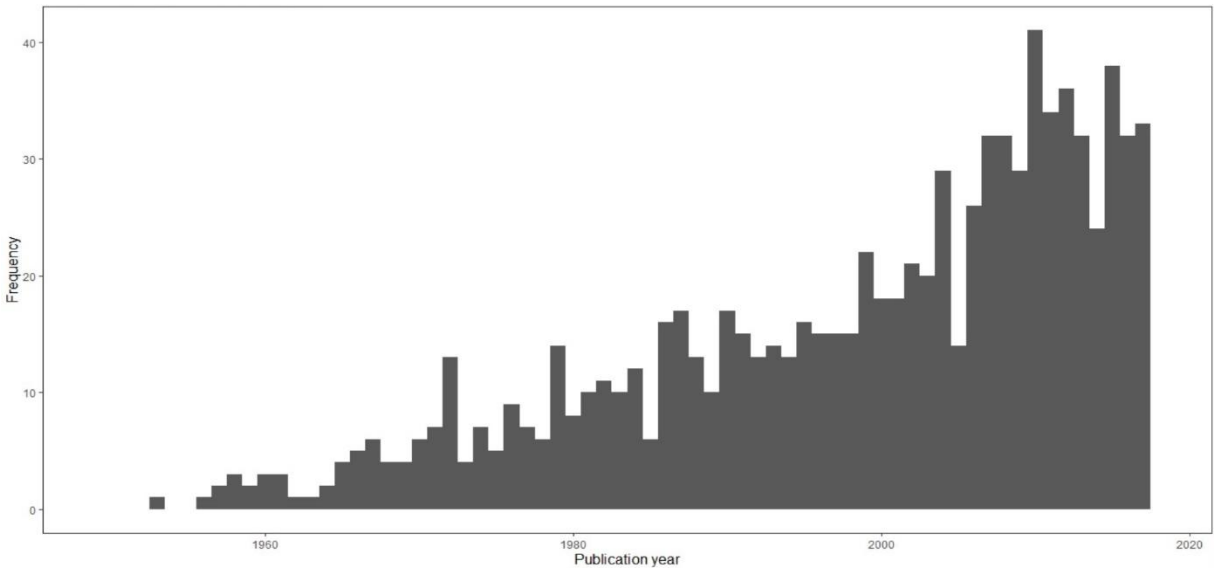


Figure 7. Distribution of all publication years in the final dataset (n = 937).

Selecting Top, Center, and Bottom 10 Studies

Of the 937 RV ranked studies in the final dataset, thirty studies are selected for further examination. First, the top 10 is defined as the ten papers with the highest RVs. Thereafter, the center 10 are obtained from the following row numbers: subtracting 4 from (937 / 2) 469 and adding 5 to 469. Next, the fifty studies with the lowest RV have a citation score of zero. These bottom fifty studies are ordered ascendingly on sample size, because a lower sample size is assumed to produce more uncertain findings (i.e., findings with a higher RV) than a higher sample size. Lastly, the ten studies from the bottom (i.e., the bottom 10) are obtained by extracting the ten studies with the highest sample size from the fifty studies with a citation score of zero.

Defining 'degrees of freedom space'. According to the RV ranking, the top 10 studies are the most valuable to be replicated and the bottom 10 are the least worthwhile to replicate. Because the top, center and bottom 10 are construed based on quantitative criteria, a replicator still has to manually go through the top papers with certain criteria in mind. In order to assess whether these top papers indeed differ from the center and bottom, in the current thesis the papers from the top, center, and bottom 10 (n = 30) are assessed on the basis of their potential research degrees of freedom. Table 2 is used for coding the papers on transparency and RDF. The first seven items on this RDF checklist are based on a literature review of QRPs and RDF (Appendix A), from which the items were selected that should be transparently reported in the original study (Dunlap, 1926). The last item of the RDF checklist concerns the construction and interpretation of a single-article *p*-curve. The *p*-curve is used to assess the evidential value of findings (Simonsohn, Simmons, and Nelson, 2015) and is part of the DFS because it indicates how likely it is that significant findings are the result of selective reporting (Simonsohn, Nelson, & Simmons, 2014). After entering statistics into the online *p*-curve app, a graph of the *p*-curve is shown accompanied by/with the results of the combination test of Simonsohn and colleagues (2015): "A set of studies is said to contain evidential value if either the half *p*-curve has a $p <$ .05 right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." (p. 1151) Besides that, "*p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) Note that the item about the *p*-curve is one of three items that can be scored zero in different ways. The other two items are about exclusion criteria and covariates, because an original study has less potential room for flexibility if no in- and exclusion criteria or covariates are used. The coding on the eight items of the RDF checklist (Table 2) was then used to create a DFS graph. Each type of RDF is formulated in such a way

that it is possible to score it solely based on the original paper (i.e. reporting completeness), and is ranked on the following suggested transparency/flexibility scale ranging from low RDF and very high transparency to high RDF and low transparency. This scale takes into account both the level of transparency of the report and the potential flexibility of the original researcher. The resulting DFS graph is thus based on a combination of reporting transparency and RDF.

Table 2

Coding on transparency and RDF

Description RDF	*0 = Low RDF / Very high transparency*	*1 = Moderate RDF / High transparency*	*2 = High RDF / Moderate transparency*	*3 = Very high RDF / Low transparency*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	Paper clearly states whether/which parts of the study were confirmatory or exploratory, and is preregistered.	Paper does not clearly state whether/which parts of the study was confirmatory or exploratory, but has some form of preregistration.	Paper clearly states whether/which parts of the study were confirmatory or exploratory, but is not preregistered.	Paper does not clearly state whether the study was confirmatory or exploratory, and is not preregistered.
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	Either: Paper does not use in- and exclusion criteria. Or: Paper clearly states beforehand which and why in- and exclusion criteria are used for selecting participants in analyses (e.g., clearly states predetermined rules about dealing with outliers).	Paper clearly states which and why in- and exclusion criteria were used for selecting participants in analyses (e.g., clearly states how outliers were dealt with).	Paper clearly states which (but not why) in- and exclusion criteria were used for selecting participants in analyses (e.g., clearly states how outliers were dealt with).	Paper does not clearly state which in- and exclusion criteria are used for selecting participants in analyses (e.g., does not clearly state how outliers were dealt with).

Sample size (predetermined or not).	Paper clearly states how the sample size or stopping rule was predetermined.	Paper clearly states that (but not how) the sample size or stopping rule was predetermined.	Paper clearly states that the sample size or stopping rule was not determined beforehand.	Paper does not clearly state whether the sample size or stopping rule was determined beforehand or not.
Sharing/ Openness (i.e., data, code, materials).	Paper shares data, code, and materials.	Paper shares two of the following: data, code, and materials.	Paper shares one of the following: data, code, and materials.	Paper shares none of the following: data, code, and materials.
Using covariates and reporting the results with and without the covariates.	Either: Paper does not use covariates. Or: Paper clearly states which and why covariates were used and the results are reported with and without the covariate(s), or only the preregistered analysis is reported.	Paper clearly states which (but not why) covariates were used and the results are reported with and without the covariate(s).	Paper clearly states which covariates were used, and the results are reported with the covariate(s). It is mentioned that the results without the covariate(s) are comparable to those with covariate(s).	Paper states that covariates were used, and the results are reported with the covariate(s), but not without the covariate(s).
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an ad hoc manner.	Paper clearly states how statistical assumptions are checked, what the outcomes were, and that violations (if any) are dealt with in a predetermined way.	Paper clearly states how statistical assumptions are checked, what the outcomes were, and how violations of statistical assumptions (if any) are dealt with.	Paper clearly states how statistical assumptions are checked, but not what the outcomes were or how violations of statistical assumptions (if any) are dealt with.	Paper does not clearly state whether statistical assumptions are checked, what the outcomes were, or how violations of statistical assumptions (if any) are dealt with.
Fallacious interpretation of (lack of)	Authors report effect sizes, and they do not	Authors report effect sizes, but they	Authors fail to report effect sizes, but they do not	Authors fail to report effect sizes and

statistical significance.	fallaciously interpret (lack of) statistical significance implying anything about the size or importance of the effect(s).	fallaciously interpret (lack of) statistical significance implying something about the size or importance of the effect(s).	fallaciously interpret (lack of) statistical significance implying something about the size or importance of the effect(s).	"authors fallaciously interpret lack of statistical significance to imply lack of effect, or weak effects may be incorrectly interpreted as important because they are statistically significant." (Rothman, 2014, p. 1063)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	Either: The paper does not disclose enough statistics to calculate the single-article *p*-curve. Or: The single-article half *p*-curve test is significantly right-skewed (i.e. $p < .05$) or both the single-article half and full *p*-curve test are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014). Furthermore, the 33% power test is $p \geq .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p \geq$	The single-article half *p*-curve test is significantly right-skewed (i.e. $p < .05$) or both the single-article half and full *p*-curve test are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014). However, the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$, which implies that the study lacks (adequate)	The single-article *p*-curve is not significantly right-skewed (i.e. $p < .05$) or both the single-article half and full *p*-curve test are not significantly right-skewed ($p < .1$), which implies that the study lacks evidential value (Simonsohn et al., 2014). However, the 33% power test is $p \geq .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p \geq .1$, which does not imply that the study lacks (adequate) evidential value (Simonsohn et al., 2015).	The single-article *p*-curve is not significantly right-skewed (i.e. $p < .05$) or both the single-article half and full *p*-curve test are not significantly right-skewed ($p < .1$), which implies that the study lacks evidential value (Simonsohn et al., 2014). Furthermore, the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$, which implies that the study lacks

.1, which does not imply that the study lacks (adequate) evidential value (Simonsohn et al., 2015).	evidential value (Simonsohn et al., 2015).	(adequate) evidential value (Simonsohn et al., 2015).

The higher the sum of the scores on the eight items from Table 2, the larger the DFS of the paper. The total of the scores can range from (8 * 0 =) 0 to (8 * 3 =) 24. The scoring on each item is visualized in a radar plot for each paper from the top, center, and bottom 10. The assumption is that the larger the area in the radar plot, the more worthwhile the study is to replicate. In what follows, the use of examining the DFS of the top ranked studies is evaluated to manually select a target for replication, by comparing the top DFSs to the center and bottom ones. Based on both the 'objective' RV formula (quantitative) and the subjective examination of DFS (qualitative), a single paper is selected from said top 10. This paper is deemed the most appropriate to be replicated for our purposes.

Results

In this section, first the distributions and frequencies of the RVs are visualized, followed by DFSs of the top, center and bottom studies. The ranking of the 937 studies on RV provided the following distribution of the RVs (*M* = .01) (Figure 8). The highest RV is approximately .536 and the smallest RVs are 0.

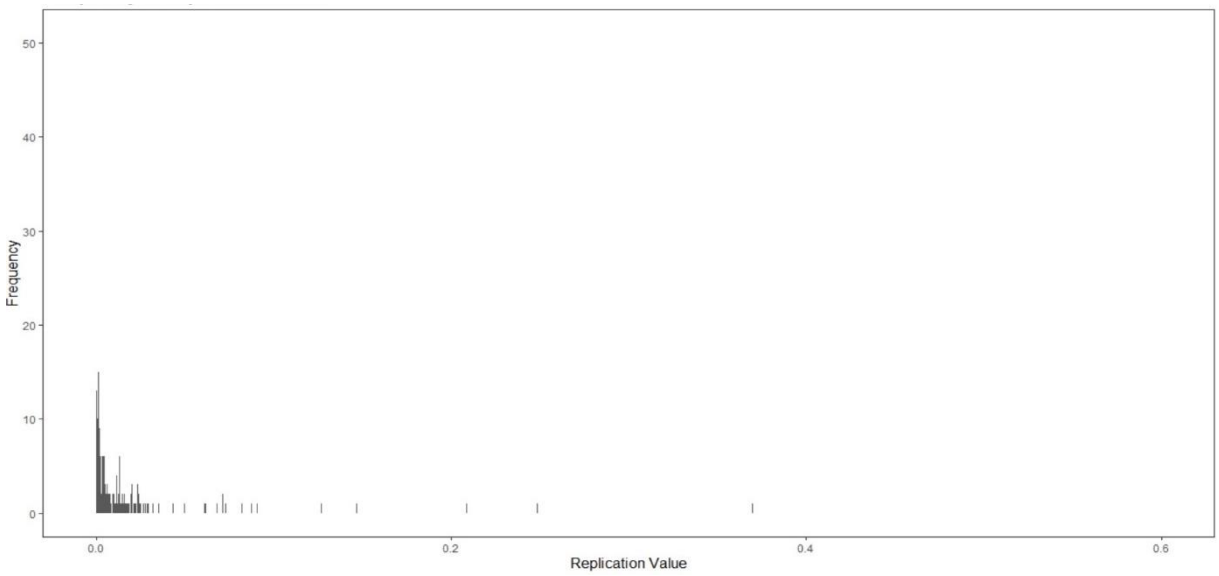


Figure 8. Distribution of all RVs in the final dataset (*n* = 937).

Top 10 Studies

In what follows, the top 10 studies with the highest RVs are examined (Table 3).

Table 3

Overview top 10 studies

Rank number	*Authors*	*Title*	*Year*	*RV*	*Citation score*	*Sample size*	*Study number*
1	Mazur, Booth, & Dabbs	Testosterone and chess competition	1992	.536	171	11	1
2	Bargh, Chaiken, Govender, & Pratto	The generality of the automatic attitude activation effect	1992	.370	633	59	3
3	Jonas, Schimel, Greenberg, & Pyszczynski	The Scrooge effect: Evidence that mortality salience increases prosocial attitudes and behavior	2002	.326	192	31	1
4	Rozin, Lowery, & Ebert	Varieties of disgust faces and the structure of disgust	1994	.260	842	120	3
5	Strack & Mussweiler	Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility	1997	.249	400	67	3
6	Mischel & Ebbesen	Attention in delay of gratification	1970	.209	341	32	1
7	Batson et al.	Is empathy-induced helping due to self-other merging?	1997	.193	278	60	2
8	Veling, Holland, & Van Knippenberg	When approach motivation and behavioral inhibition collide: Behavior regulation through stimulus devaluation	2008	.186	80	33	1
9	Lodge & Taber	The automaticity of affect for political leaders, groups, and issues: An experimental test of	2005	.167	214	80	1

		the hot cognition hypothesis					
10	Stellar, Cohen, Oveis, & Keltner	Affective and physiological responses to the suffering of others: Compassion and vagal activity	2015	.147	45	51	1

In order to map the DFS of the top 10 studies, each of the ten studies was scored on RDF (see Appendix B for an elaboration on how each study is scored). The scores are summarized in Table 4. Recall that the scores per item range from 0 (lowest RDF/highest transparency) to 3 (highest RDF/lowest transparency).

Table 4

Scoring the top 10 studies

RDF	Nr. 1	Nr. 2	Nr. 3	Nr. 4	Nr. 5	Nr. 6	Nr. 7	Nr. 8	Nr. 9	Nr. 10
Confirmatory vs. exploratory	2	3	2	3	2	2	2	2	2	2
Exclusion of participants	1	0	0	3	1	1	1	0	1	3
Sample size	2	3	3	3	3	3	3	3	3	3
Sharing/Openness	3	2	2	2	2	2	2	3	2	3
Covariates	0	0	0	0	0	0	3	0	0	0
Statistical assumptions	3	3	3	3	3	3	3	3	3	3
Effect sizes	3	1	3	2	3	3	2	0	3	0
Single-article *p*-curve	0	0	3	0	0	2	0	3	0	0
Total score	14	12	16	16	14	16	16	14	14	14

RDF Patterns in Top 10

In several ways, the top 10 studies differ in their scores on the eight items of the RDF checklist (Table 2). The scores on the item about the transparency of reporting about which participants are excluded and why, vary a lot between the top 10 studies. Note that the three studies (i.e., nr. 2, nr. 3, and nr. 8) that scored zero on this item, all did not apply any in- or

exclusion criteria. Thus, it is not the case that they scored zero, because they clearly stated beforehand which and why in- and exclusion criteria were used. Another item which had fluctuating scores between the top 10 studies, is the item about effect sizes and interpreting statistical significance. Three studies (i.e., nr. 2, nr. 8, and nr. 10) reported effect sizes, while the remaining seven failed to do so. Half of the top 10 studies did not only fail to report effect sizes but also fallaciously interpreted (lack of) statistical significance. The single-article *p*-curves also varied between the top 10 studies. Nr. 3 and nr. 8 seemed to lack (adequate) evidential value. The seven studies that scored the best on this item (i.e., a score of zero) can be split into two groups: whereas two studies (i.e., nr. 1 and nr. 4) did not provide enough information to put in the *p*-curve app, the remaining five studies (i.e., nr. 2, nr. 5, nr. 7, nr. 9, and nr. 10) generated a *p*-curve that indicates (adequate) evidential value. Furthermore, the total flexibility scores of the top 10 studies range from 12 to 16. It is noteworthy that the highest score (i.e., 16) belongs to four of the ten studies. Moreover, the nr. 1 study scored 14 in total and the nr. 2 scored 12. Both did not have the highest total flexibility/transparency score. This may indicate that the qualitative analysis is a useful addition to accompany the RV formula.

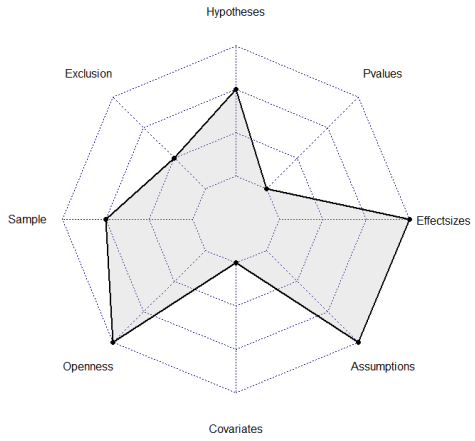
In other ways, the top 10 studies are similar in their scores on the eight RDF items (Table 2). The most striking similarity is that all but one of the top 10 studies did not contain covariates and therefore got assigned 0 points on said item. Thus, the reason for scoring zero was not that the paper clearly states which and why covariates were used and that the results are reported with and without the covariate(s). Another item that scored quite similar between the top 10 studies is whether it is clearly stated if the study is confirmatory or exploratory. The reason that none of the top 10 studies scored 0 or 1 on this item, is that none of them are (partially) preregistered. The same goes for the lack of predetermined sample sizes or stopping rules. Likewise, all of the top 10 studies did not share any or just one of the following: data, code, and materials.

DFS Graphs of Top 10

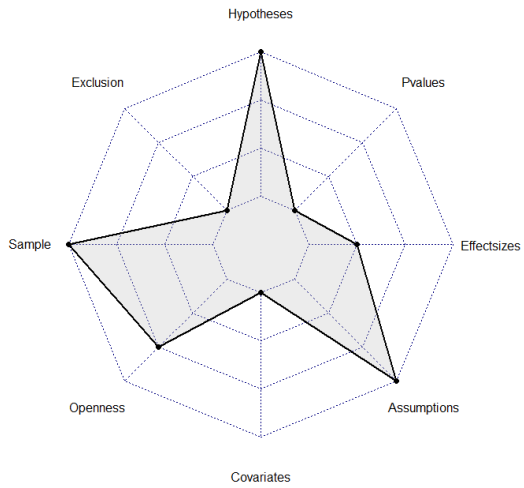
The DFS graphs for the top 10 studies (Figure 9a and 9b) are constructed based on the scores on the RDF checklist. Each dotted octagon within the graph represents the scores 0, 1, 2, and 3 respectively. For example, if a study scores the highest possible score of 3 on an item, this is mapped as a line to the outer edge of the graph. The area of each graph of the top 10 is fairly large, especially nr. 4 with the highest total score of 19. Nr. 8 stands out, because after running this *p*-curve in the app ("*P*-curve app 4.06," 2017), the following comment appeared in bold text: "direct replications of the submitted studies are not expected to succeed." ("*P*-

curve results app 4.06," 2017) This comment was nowhere to be found after running the *p*-curves for any of the other studies in the top 10.

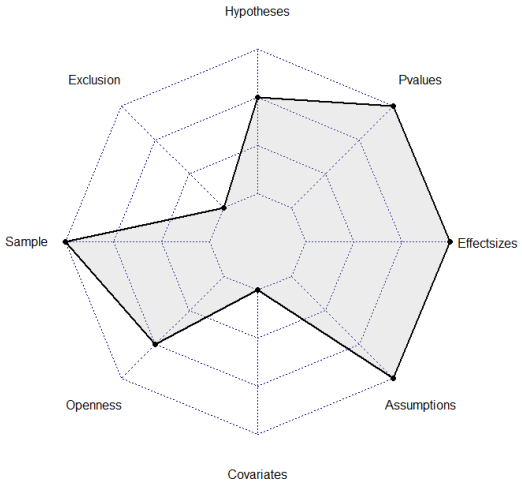
1: Testosterone and Chess Competition



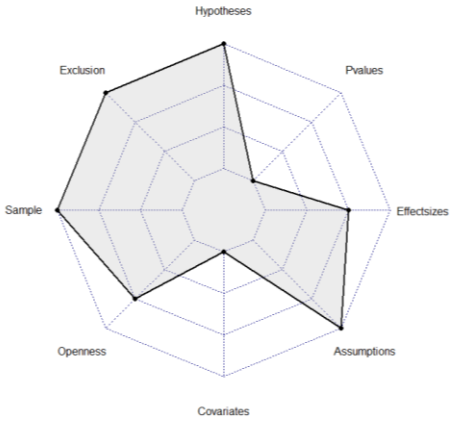
2: Generality of the Automatic Attitude Activation Effect



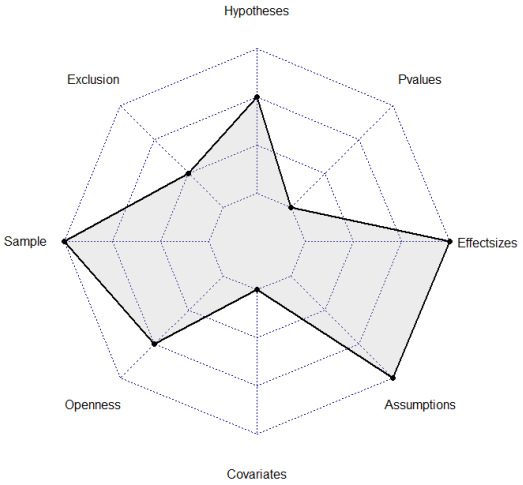
3: The Scrooge Effect: Evidence That Mortality Salience Increases Prosocial Attitudes and Behavior



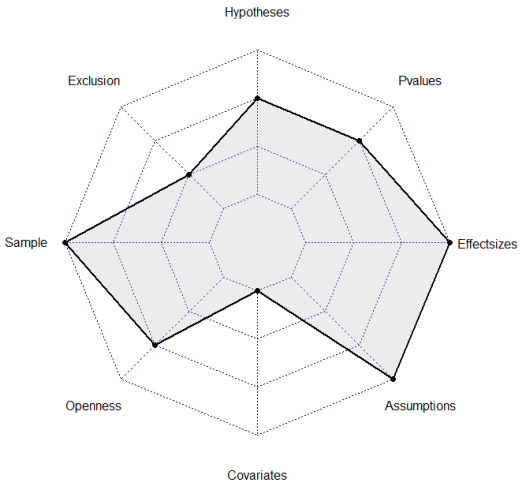
4: Varieties of Disgust Faces and the Structure of Disgust



5: Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility



6: Attention in Delay of Gratification



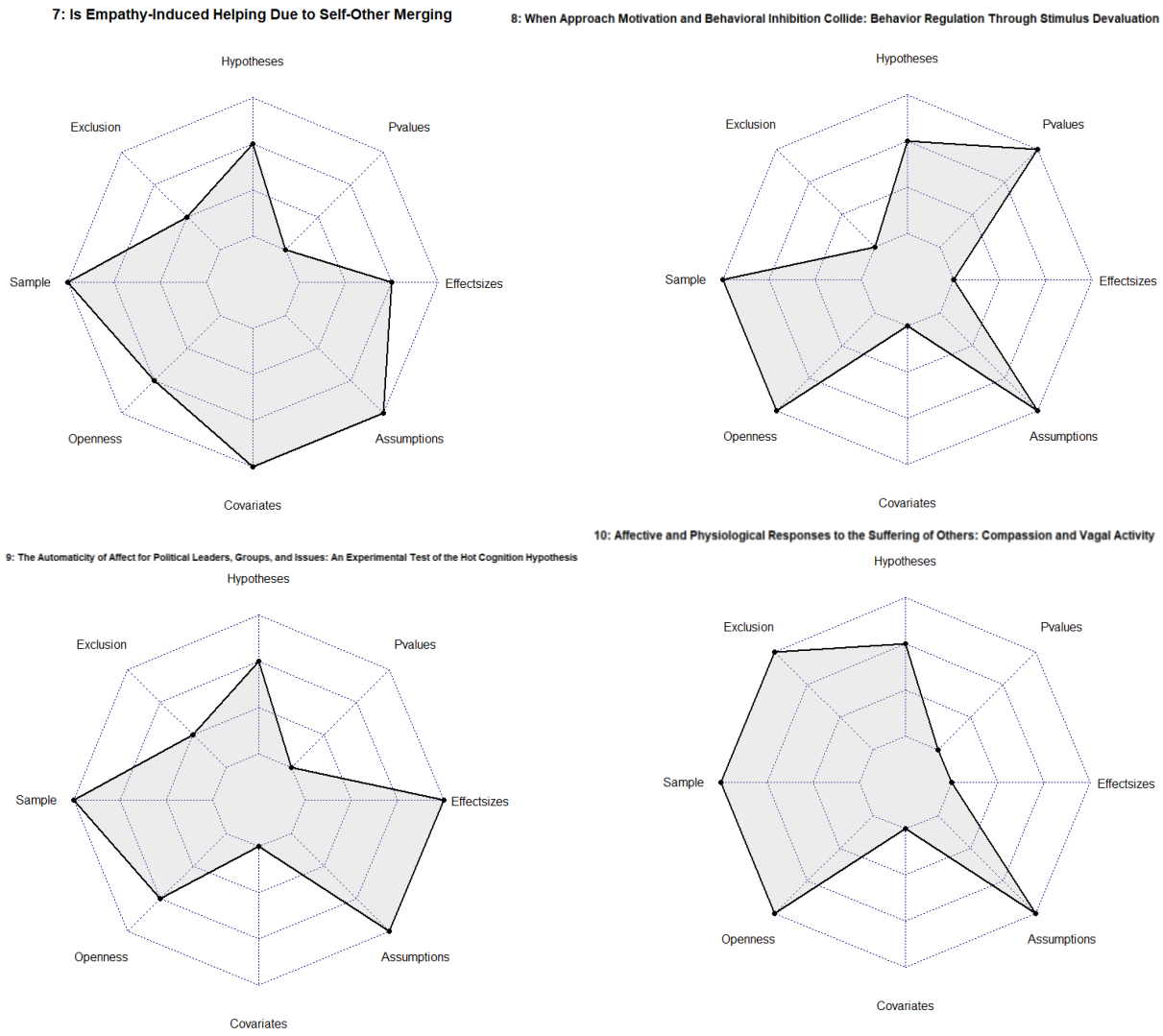


Figure 9a. Individual radar charts of all studies in the top 10.

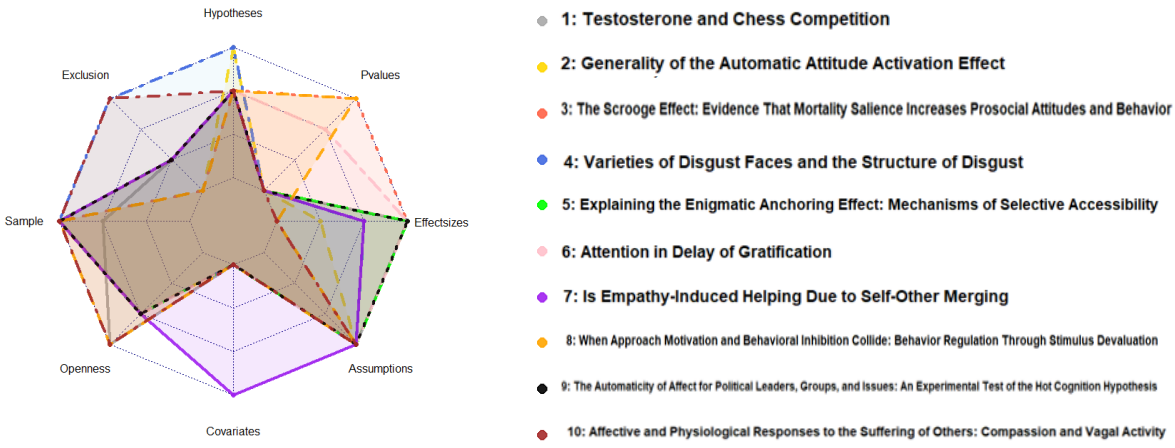


Figure 9b. Merged radar chart of all studies in the top 10.

Center 10 Studies

In what follows, the center 10 studies with the middle/median RVs are examined (Table 5).

Table 5

Overview center 10 studies

Rank number	*Authors*	*Title*	*Year*	*RV*	*Citation score*	*Sample size*	*Study number*
1	Griese, McMahon, & Kenyon	A research experience for American Indian undergraduates: Utilizing an actor–partner interdependence model to examine the student–mentor dyad	2017	.005	1	53	1
2	Bromet & Moos	Environmental Resources and the Posttreatment Functioning of Alcoholic Patients	1977	.005	89	429	1
3	Berant & Wald	Self-reported attachment patterns and Rorschach-related scores of ego boundary, defensive processes, and thinking disorders	2009	.005	5	89	1
4	Morrison	A license to speak up: Outgroup minorities and opinion expression	2011	.005	8	172	2
5	Galanis & Jones	When stigma confronts stigma: Some conditions enhancing a victim's tolerance of other victims	1986	.005	13	80	1
6	Hevey et al.	Consideration of future consequences scale: Confirmatory Factor Analysis	2010	.005	30	590	1

7	Cameron	Social identity, modern sexism, and perceptions of personal and group discrimination by women and men	2001	.005	28	303	1
8	Gyurcsik, Brawley, & Langhout	Acute thoughts, exercise consistency, and coping self-efficacy	2002	.005	14	160	1
9	Thieme & Feij	Tyramine, a new clue to disinhibition and sensation seeking?	1986	.005	4	25	1
10	Surmann	The effects of race, weight, and gender on evaluations of writing competence	1997	.005	7	64	1

In order to map the DFS of the top 10 studies, each of the ten studies is scored on RDF (see Appendix C for the elaboration on how each study is scored). The scores are summarized in Table 6. Recall that the scores per item range from 0 (lowest RDF/highest transparency) to 3 (highest RDF/lowest transparency).

Table 6

Scoring the center 10 studies

RDF	Nr. 1	Nr. 2	Nr. 3	Nr. 4	Nr. 5	Nr. 6	Nr. 7	Nr. 8	Nr. 9	Nr. 10
Confirmatory vs. exploratory	2	2	2	2	2	2	2	2	2	3
Exclusion of participants	0	1	0	1	0	0	0	3	1	0
Sample size	3	3	3	3	3	2	3	2	3	3
Sharing/Openness	3	2	3	2	2	2	2	2	2	3
Covariates	0	0	0	0	0	0	3	3	0	3
Statistical assumptions	1	1	1	3	1	3	3	1	3	3
Effect sizes	2	2	2	2	3	2	2	0	2	3
Single-article *p*-curve	0	0	0	2	2	0	2	0	0	2

Total score	11	11	11	15	13	11	17	13	13	20

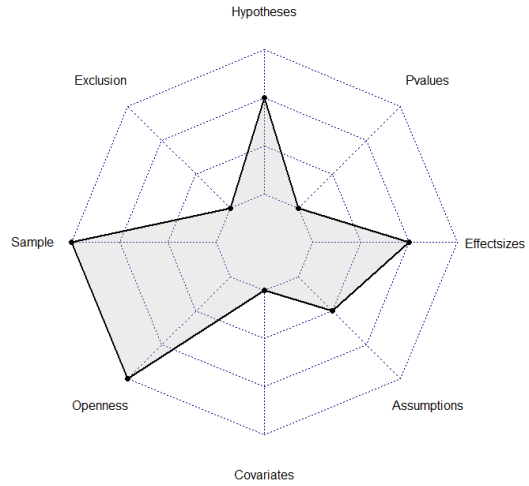
RDF Patterns in Center 10

In several ways, the center 10 studies differ in their scores on the eight items of the RDF checklist (Table 6). The scores on the item about the transparency of reporting about which participants are excluded and why differ a bit between the center 10 studies. Note that the six studies that scored zero on this item, all did not apply any in- or exclusion criteria. Thus, it is not the case that they scored zero, because they clearly stated beforehand which and why in- and exclusion criteria were used. Another item which had fluctuating scores between the center 10 studies, is the item about checking the statistical assumptions: studies in the center 10 either clearly stated how they were checked and what the outcomes were, or they did not. Furthermore, the total flexibility/transparency scores of the center 10 studies range from 11 to 20. It is noteworthy that the lowest score (i.e., 20) belongs to the nr. 10 study and not to the nr. 1. Based on the RV formula alone, nr. 1 would be expected to be the least worthwhile to replicate (i.e., have the lowest score) out of these ten studies. This may indicate that the qualitative analysis is a useful addition to accompany the RV formula.

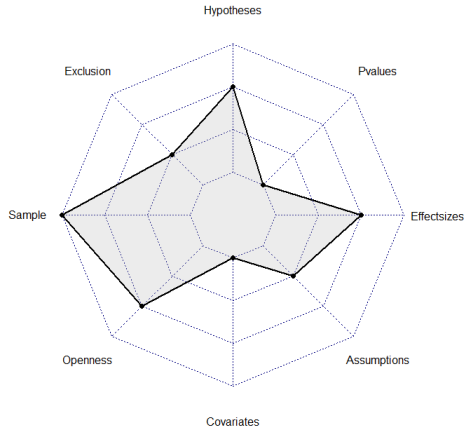
In other ways, the center 10 studies are similar in their scores on the eight RDF items (Table 6). The most striking similarity is that all but one of the center 10 studies clearly stated whether/which parts of the study were confirmatory or exploratory without being preregistered. Furthermore, seven out of the ten studies got assigned zero points on the item about covariates: six of them because they did not contain covariates, but nr. 9 scored zero because the results were reported both with and without the covariates. Two other items that scored quite similar between the center 10 studies are those about sample size and openness.

DFS Graphs of Center 10

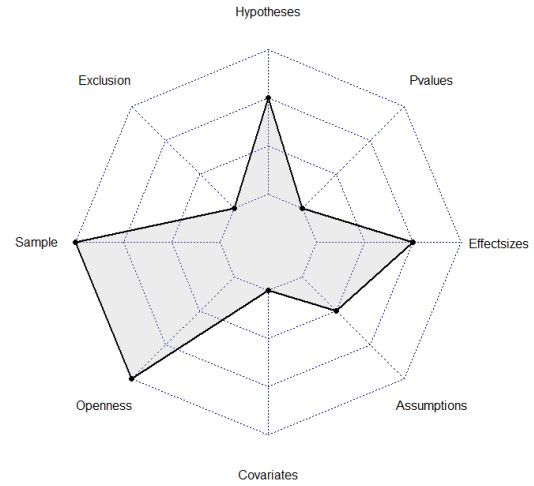
The DFS graphs for the center 10 studies (Figure 10a and 10b) are constructed based on the scores on the RDF checklist. Each dotted octagon within the graph represents the scores 0, 1, 2, and 3 respectively. For example, if a study scores the highest possible score of 3 on an item, this is mapped as a line to the outer edge of the graph. The areas of the graph differ a lot in size.



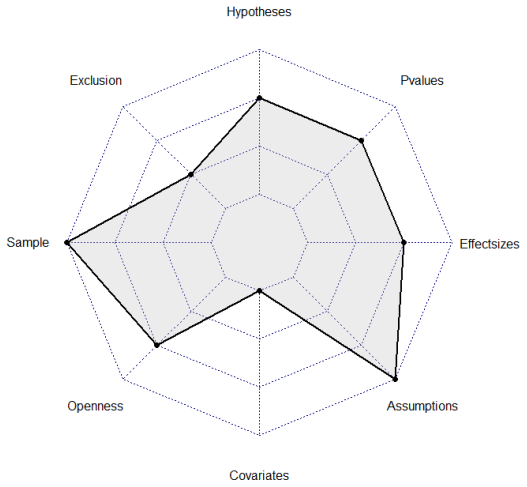
2: Environmental Resources and the Posttreatment Functioning of Alcoholic Patients



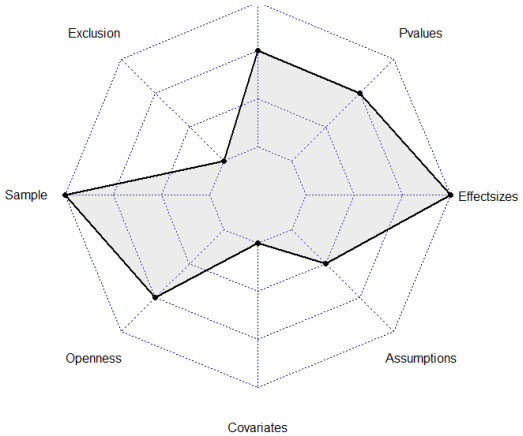
3: Self-Reported Attachment Patterns and Rorschach-Related Scores of Ego Boundary, Defensive Processes, and Thinking Disorders



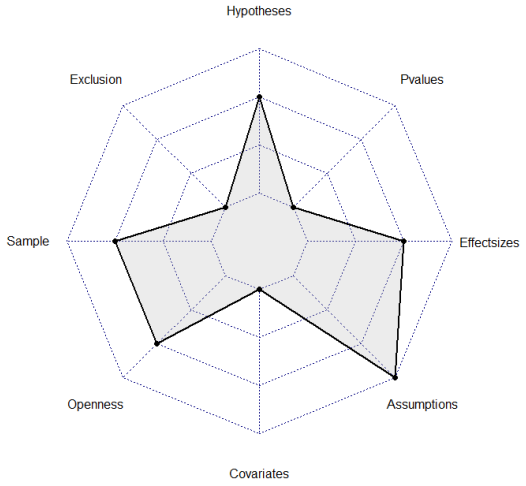
4: A license to speak up: Outgroup minorities and opinion expression



5: When Stigma Confronts Stigma: Some Conditions Enhancing a Victim's Tolerance of Other Victims



6: Consideration of future consequences scale: Confirmatory Factor Analysis



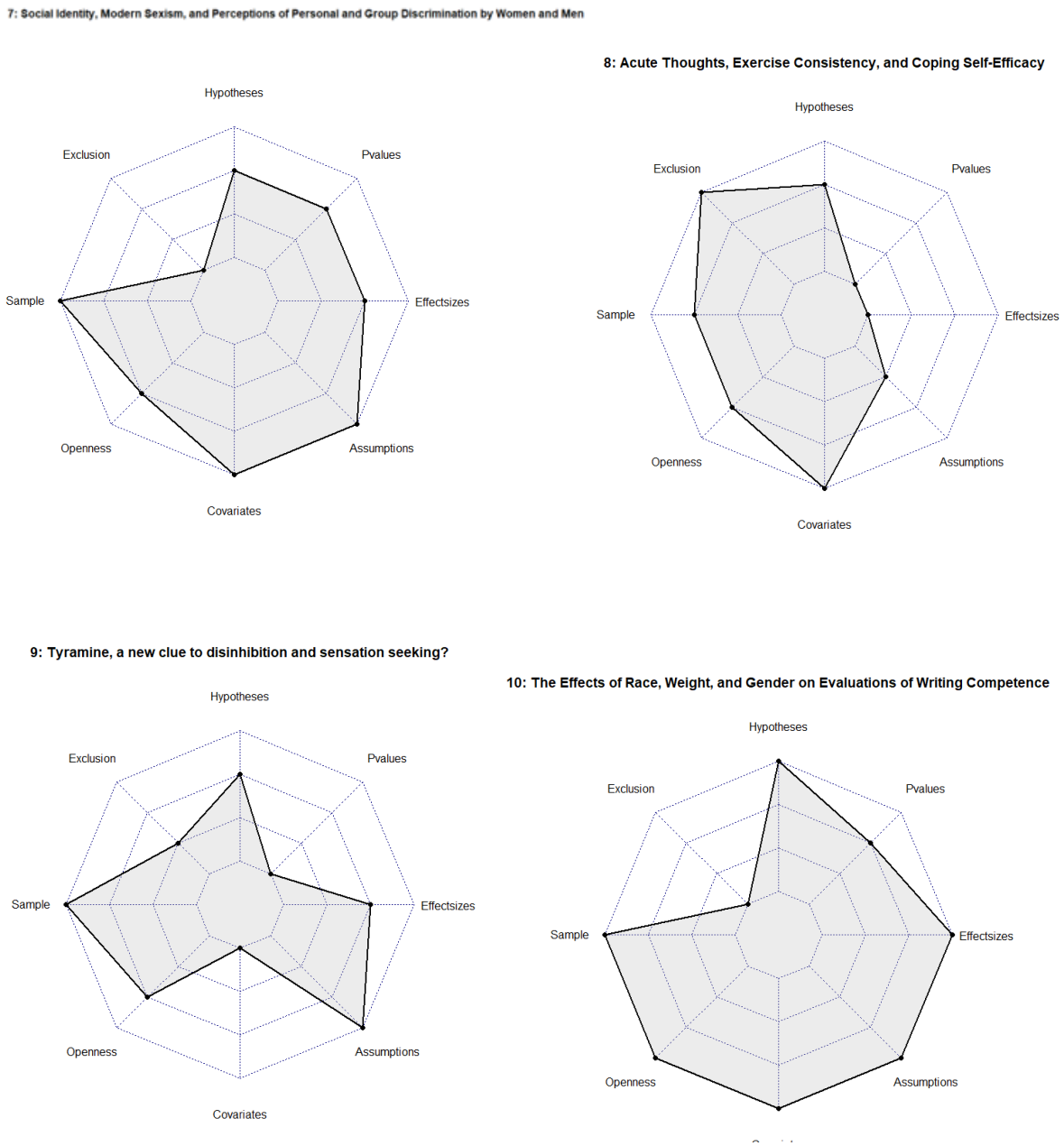


Figure 10a. Individual radar charts of all studies in the center 10.

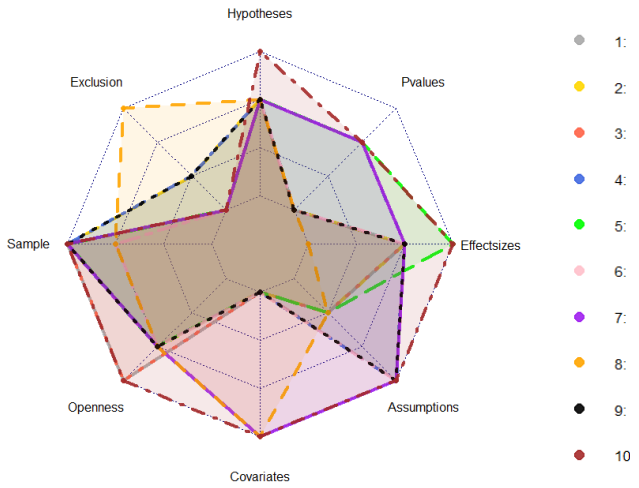


Figure 10b. Merged radar chart of all studies in the center 10.

Bottom 10 Studies

In what follows, the bottom 10 studies with the lowest RVs are examined (Table 7).

Table 7

Overview bottom 10 studies

Rank number	*Authors*	*Title*	*Year*	*RV*	*Citation score*	*Sample size*	*Study number*
1	Puddifoot	The persuasive effects of a real and complex communication	1996	0	0	3713	1
2	Brundidge, Baek, Johnson, & Williams	Does the medium still matter? The influence of gender and political connectedness on contacting U.S. public officials online and offline	2013	0	0	2251	1
3	Silva, Delerue Matos, & Martinez-Pecino	Confidant network and quality of life of individuals aged 50+: the positive role of internet use	2018	0	0	1828	1
4	Oyamot, Jackson, Fisher,	Social norms and egalitarian values mitigate authoritarian	2017	0	0	1212	1

	Deason, & Borgida	intolerance toward sexual minorities					
5	Santens et al.	Personality profiles in substance use disorders: Do they differ in clinical symptomatology, personality disorders and coping?	2018	0	0	700	1
6	Peacock, Cowan, Bommersbach, Smith, & Stahly	Pretrial predictors of judgments in the O.J. Simpson case	1997	0	0	578	1
7	Kalibatseva & Leong	Cultural factors, depressive and somatic symptoms among Chinese American and European American college students	2018	0	0	519	1
8	Burtăverde, De Raad, & Zanfirescu	An emic-etic approach to personality assessment in predicting social adaptation, risky social behaviors, status striving and social affirmation	2018	0	0	515	1
9	Thomas & Mucherah	Brazilian adolescents' just world beliefs and its relationships with school fairness, student conduct, and legal authorities	2018	0	0	475	1
10	Zhang, Qiu, & Teng	Cross-level relationships between justice climate and organizational citizenship behavior: Perceived organizational support as mediator	2017	0	0	468	1

In order to map the DFS of the top 10 studies, each of the ten studies is scored on RDF (see Appendix D for the elaboration on how each study is scored). The scores are summarized in Table 8. Recall that the scores per item range from 0 (lowest RDF/highest transparency) to 3 (highest RDF/lowest transparency).

Table 8

Scoring the bottom 10 studies

RDF	Nr. 1	Nr. 2	Nr. 3	Nr. 4	Nr. 5	Nr. 6	Nr. 7	Nr. 8	Nr. 9	Nr. 10
Confirmatory vs. exploratory	2	2	2	2	2	2	2	2	2	2
Exclusion of participants	0	1	0	0	2	0	0	2	0	0
Sample size	1	3	3	0	3	3	3	3	3	3
Sharing/Openness	2	2	3	1	2	2	2	2	3	3
Covariates	0	3	3	0	0	0	0	0	2	3
Statistical assumptions	3	3	1	3	3	3	3	3	1	2
Effect sizes	2	0	0	0	2	2	0	2	3	3
Single-article *p*-curve	0	0	0	0	0	0	0	0	0	0
Total score	10	14	12	6	14	12	10	14	14	16

RDF Patterns in Bottom 10

In several ways, the bottom 10 studies differ in their scores on the eight items of the RDF checklist (Table 2). The scores on the item about openness fluctuated between sharing zero, one, or two of the following: data, code, and materials. None of the studies shared all three. Furthermore, four studies (i.e., nr. 2, nr. 3, nr. 4, and nr. 7) did not only report effect sizes but also interpreted (lack of) statistical significance correctly (i.e., as not implying anything about the size of importance of the effect(s)). Four other studies (i.e., nr. 1, nr. 5, nr. 6, and nr. 8) failed to report effect sizes, but they also did not misinterpret (lack of) statistical significance. The majority of the studies (i.e., seven out of ten) did not report clearly whether statistical assumptions are checked, what the outcomes of those checks were, or how violations of statistical assumptions (if any) are dealt with. Another majority of the studies (i.e., seven out of

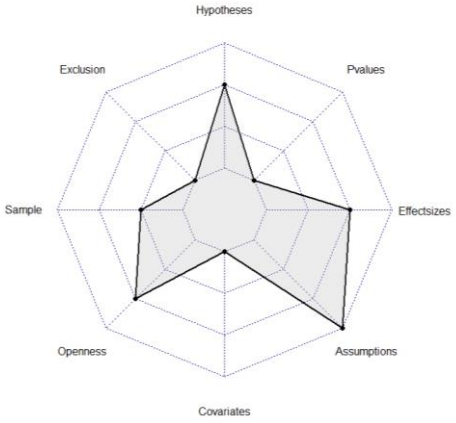
ten) scored zero on the item about exclusion criteria, because they did not contain any in- or exclusion criteria. Thus, the reason for scoring zero was not that the paper clearly stated beforehand which and why in- and exclusion criteria were used for selecting participants in analyses. Furthermore, the total flexibility/transparency scores of the bottom 10 studies range from 6 to 16. It is noteworthy that the lowest score (i.e., 6) belongs to the nr. 4 study and not to the nr. 1. Based on the RV formula alone, nr. 1 would be expected to be the least worthwhile to replicate (i.e., have the lowest score). This may indicate that the qualitative analysis is a useful addition to accompany the RV formula.

In other ways, the bottom 10 studies are similar in their scores on the eight RDF items (Table 2). The most striking similarity is that all bottom 10 studies clearly stated whether/which parts of the study were confirmatory or exploratory, but none of them was preregistered. Another noteworthy similarity is that all bottom 10 studies scored the best possible (i.e., 0) on the item about *p*-curves. However, they did differ in the reason why they scored 0: although three studies (i.e., nr. 1, nr. 3, and nr. 8) did not disclose enough statistics to calculate the *p*-curve, the remaining seven studies generated a *p*-curve that indicates (adequate) evidential value. Besides that, all of the six studies that scored zero on the item about covariates (i.e., nr. 1, nr. 4, nr. 5, nr. 6, nr. 7, and nr. 8) did not contain any covariates. Thus, the reason for scoring zero was not that the paper clearly states which and why covariates were used and that the results are reported with and without the covariate(s). Furthermore, for eight out of the bottom 10 studies, it was unclear whether the sample size or stopping rule was determined beforehand or not.

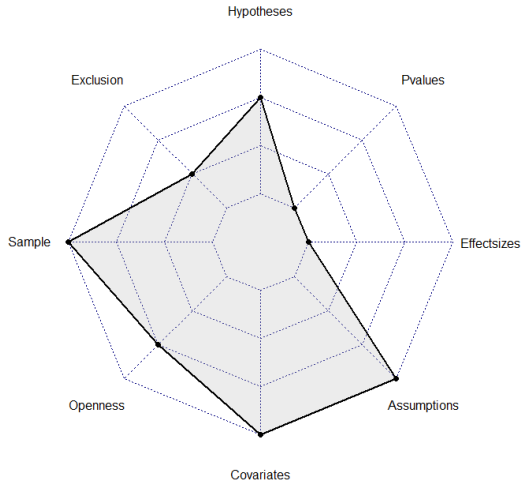
DFS Graphs of Bottom 10

The DFS graphs for the bottom 10 studies (Figure 11a and 11b) are constructed based on these scores. Each dotted octagon within the graph represents the scores 0, 1, 2, and 3 respectively. For example, if a study scores the highest possible score of 3 on an item, this is mapped as a line to the outer edge of the graph. It is noteworthy that number 2 (Brundidge et al., 2013) of the bottom 10 is a replication study. It makes sense that a replication study is not worthwhile to replicate.

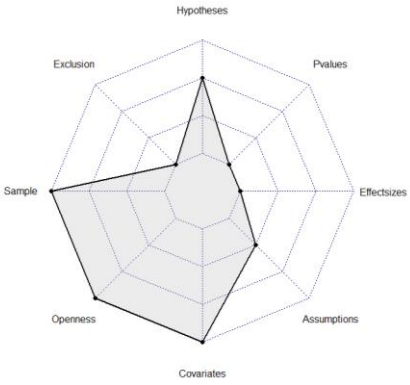
1: The Persuasive Effects of a Real and Complex Communication



2: Does the Medium still Matter? The Influence of Gender and Political Connectedness on Contacting U.S. Public Officials Online and Offline

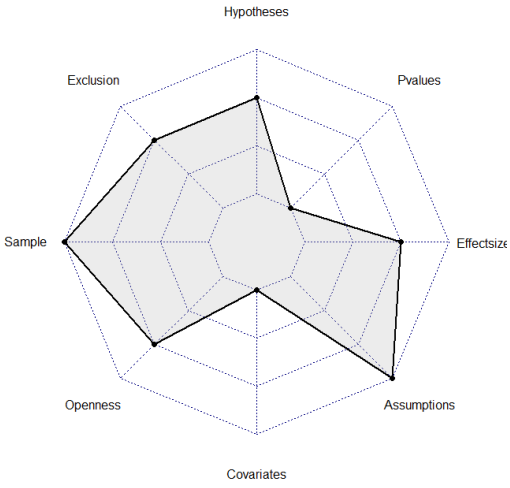


3: Confidant Network and Quality of Life of Individuals Aged 50+: The Positive Role of Internet Use

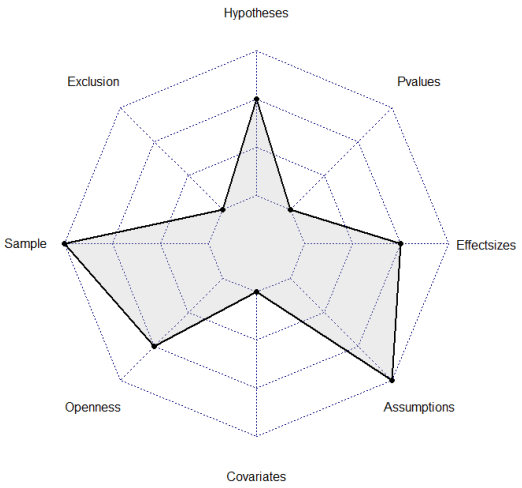


4: Social Norms and Egalitarian Values Mitigate Authoritarian Intolerance Toward Sexual Minorities

5: Personality profiles in substance use disorders: Do they differ in clinical symptomatology, personality disorders and coping?



6: Pretrial Predictors of Judgments in the O. J. Simpson Case



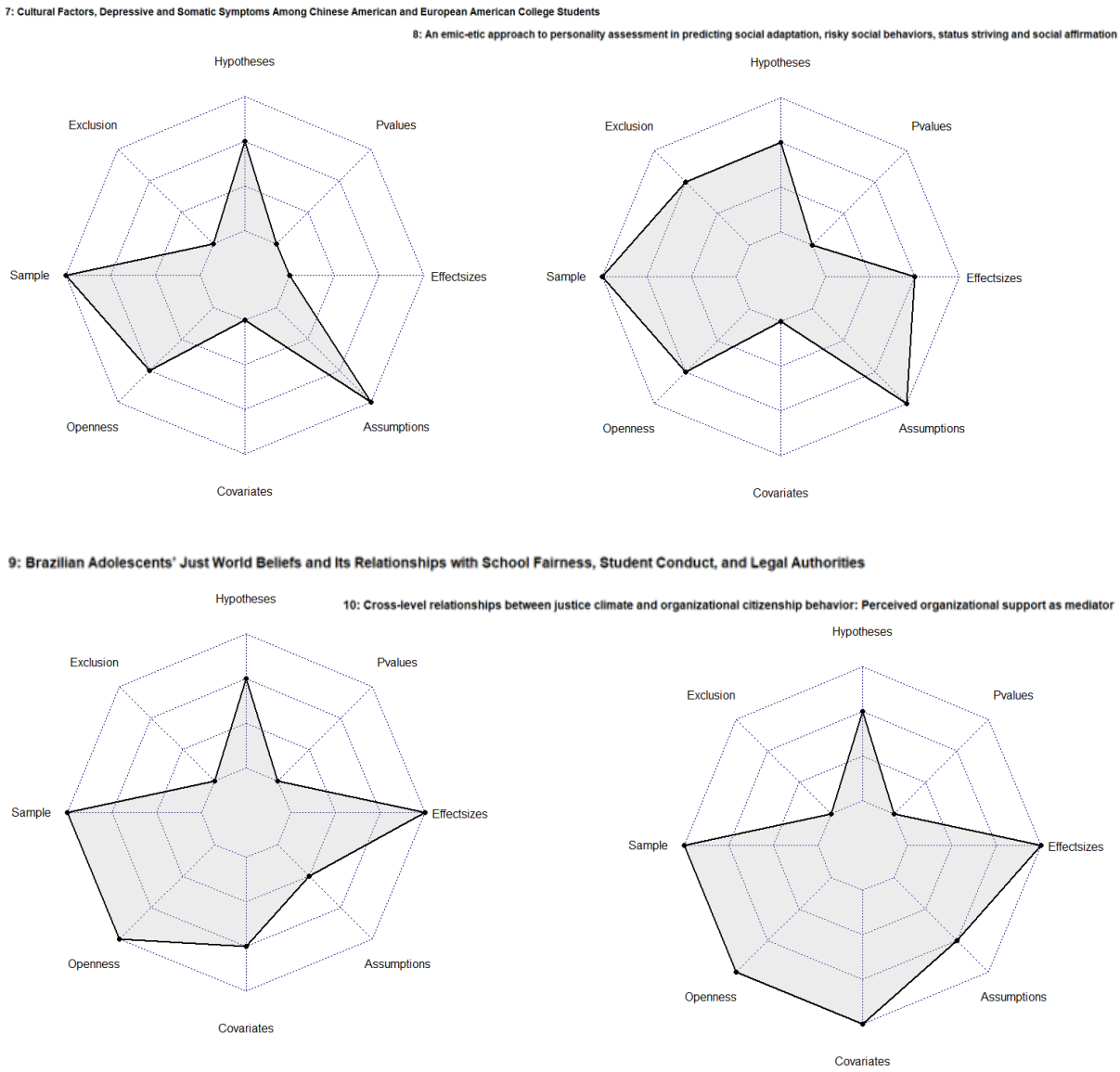


Figure 11a. Individual radar charts of all studies in the bottom 10.

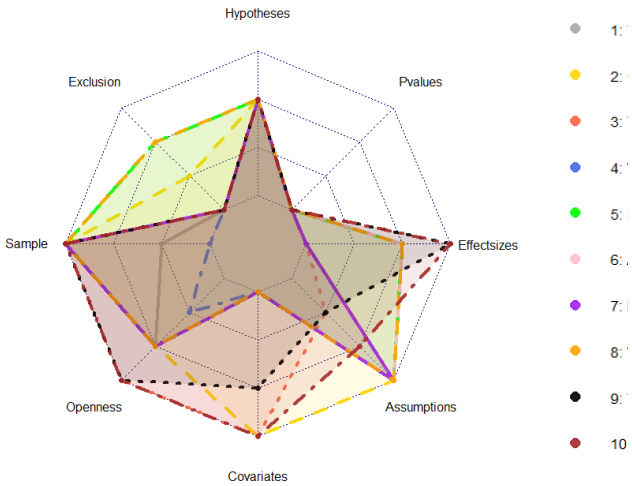


Figure 11b. Merged radar chart of all studies in the bottom 10.

Comparison of the Top, Center, and Bottom

Overall, the top 10 studies had a larger DFS than both the center and bottom 10 studies. However, it is striking that the two studies with the highest overall flexibility/transparency score belong to the center 10, and not to the top 10. Furthermore, none of the 30 papers had (some form of) preregistration, nor did any of them share data, code, and materials. The mean of the total flexibility/transparency scores is 14.6 for the top 10 studies, 13.5 for the center 10, and 12.2 for the bottom 10. The difference between the top and bottom 10 may be related to sample size or citation score, and since the bottom 10 studies have smaller DFS maps, this is aligned with RV ranking on uncertainty. As can be seen in the merged radar charts (Figure 12), this pattern also can be seen in the DFS maps. The top 10 DFSs are the largest and the bottom 10 DFSs are the smallest. The center 10 DFSs fall in between those of the top and bottom 10.

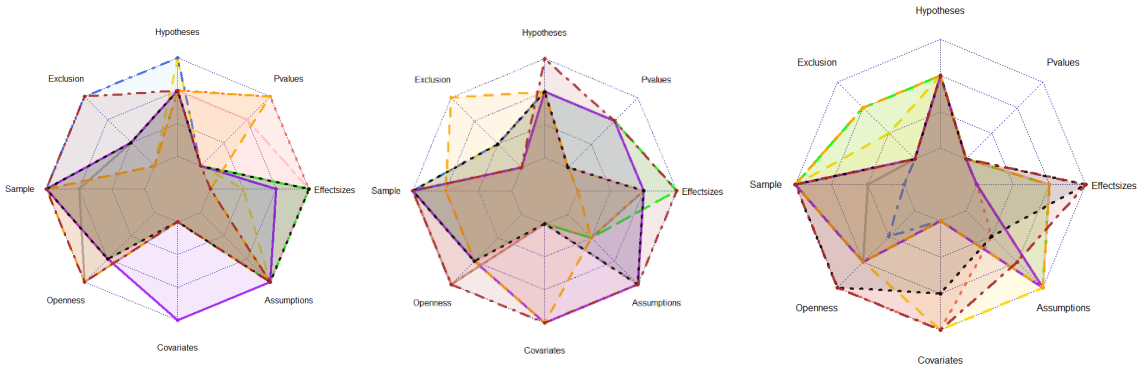


Figure 12. Merged radar chart of all studies in the top 10 (left), center 10 (middle), and bottom 10 (right).

Conclusion

The present work commenced with posing two exploratory questions. The first question is: What are the characteristics (both similarities and differences) of the top 10 of a ranking of studies based on a Replication Value (which is in turn based on sample size and citation score) compared with center and bottom ranked studies when looking at the ‘degrees of freedom space’ of the original work? The second question is: Can certain characteristics be used to map the original researcher’s ‘degrees of freedom space’ – an indicator of the reproducibility of the decisions made by the researcher(s) – in order to aid in the selection of the paper that is most worthy of replication? Based on the qualitative analysis, it seems that the RV formula did manage to rank the studies according to their need to be replicated. The top 10 studies are deemed more worthwhile to replicate than the center and bottom ranked studies after analyzing them with a focus on the DFS of the original researcher(s) (i.e., reading the top 10 studies and determining whether the DFS is bigger compared to the center 10 and bottom 10). The DFS

graphs of the bottom 10 studies had smaller areas than those of the top 10 studies. This implies that the bottom 10 studies were more transparent in reporting and had less room for flexibility than the top 10. All in all, it seems that the original authors of the top 10 studies are reporting less transparently and have more potential for exploiting their RDF compared to both the center and bottom 10 studies. This result suggests that the mapping the DFS of studies is helpful in selecting which study to replicate after making a larger selection based on RV.

Discussion

It is unclear whether aforementioned conclusions are applicable to other fields than Social Psychology. Given the exploratory nature of the current thesis, it is important to address several limitations in order to take these into account in future research concerning the topic(s) of the current research. One of those limitations is that the top, center, and bottom 10 studies are coded by a single researcher. Given the subjective nature of coding the items on the RDF checklist, it is recommended that this is done by several coders in future research and that their interrater reliability is checked. However, this was not possible for the current thesis as a consequence of operating under resource constraints.

Moreover, the DFS is based on transparency of the written reports about studies. The underlying assumption is that if the reporting is not transparent, it cannot be said with certainty whether the original researchers truthfully reported about the details of their studies. Therefore, judging the reporting completeness is as good of a measure as it gets.

Furthermore, it is unclear how representative the random sample (n = 999) of the entire field of social psychology. Although assessing whether the usefulness of DFS can also be done based on a non-representative sample, a representative sample is preferred for researchers planning to actually replicate a study after mapping the DFSs.

Combining the results from the current thesis and its limitations, it can be argued that qualitatively checking a subset of the RV ranked studies is a useful addition to effectively determine which studies should actually be replicated in practice. Researchers looking to select a target for replication can use the RDF checklist and the resulting DFS graphs to map the transparency in reporting regarding the room for flexibility of original researchers. As it is likely that more uncertainty surrounding the original work will increase the difficulty of replicating said work, it is recommended that the feasibility of replicating a study is also taken into account when selecting a target for replication. Furthermore, future research might look into ways to account for the practical impact of findings instead of only the academic impact.

Replication: the way forward

The (added) value of the current thesis can be challenged. In general, there are opposing voices speaking out against replication. For example, it could be pointed out that constraining decision flexibility will not account for false positives resulting from fairly obtained significant *p*-values. However, with a significance level of alpha =. 05 as threshold, there is always 5% chance of randomness causing false positives.

In addition, it could be mentioned that a single failed replication attempt could be due to random error or unknown variables (Crandall & Sherman, 2016). Although it is true that a single failed attempt at replicating does not necessarily invalidate the original findings (Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016), replication still is beneficial for the precision, accuracy, and veracity of findings (Ebersole et al., 2016).

Moreover, it could be emphasized that replications cost time, effort, and/or money to perform. However, it simply just takes some time getting used to (Baker, 2016). Besides this, the costs of not doing replications is arguably larger: Without replications, discovered effects are either genuine but not confirmed, or ungenuine but not challenged (Ioannidis, 2012). Scientific claims that are not (yet) replicated might as well result from random error and/or bias (Ioannidis, 2012).

Furthermore, it could be suggested that it is harder to stop researchers from exploiting their degrees of freedom than it is to reward researchers who do not exploit their researcher degrees of freedom. However, replication can be seen as a way to reward replicable studies with true positives. As pointed out by Block and Kuckertz (2018), replications are necessary "to develop convincing, robust, and reliable structured literature reviews and quantitative meta-analyses" (p. 356).

While recognizing that replication might have downsides and that certain difficulties (e.g., deciding whether to conceptually or directly replicate, and contacting the original authors) have to be overcome when replicating a study, it is argued that the benefits of replication outweigh its shortcomings. Replication is necessary going forward to repair the damaged self-correcting functions of science (Ioannidis, 2012). Therefore, it is useful to replicate studies with the most room for flexibility, because those are most likely to be false positives. The RV formula is an efficient way to quantitatively filter through an entire field and the RDF checklist can then be used to qualitatively assess the replication value of a much smaller set of studies. Although the current thesis is by no means enough to establish the RV formula combined with the RDF checklist as a robust way to aid researchers in selecting which studies to replicate, it

is a step in the direction of examining and exploring different methods that can ultimately facilitate more replication research.

References

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for Psychology* (6th ed.). New York City, NY: Pearson Education.

Artino Jr., A. R. (2013). Why don't we conduct replication studies in medical education? *Medical Education*, *47*(7), 746-747. https://doi.org/10.1111/medu.12204

Baker, M. (2016). Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and they think will help. *Nature*, *533*(7604), 452-454.

Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., & Rogelberg, S. G. (2019). Answers to 18 questions about open science practices. *Journal of Business and Psychology*, *34*, 257-270. https://doi.org/10.1007/s10869-018-9547-8

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology*, *62*(6), 893-912. https://doi.org/10.1037/0022-3514.62.6.893

Bastow, S., Dunleavy, P., & Tinkler, J. (2014). *The impact of the social sciences.* London: SAGE Publications.

Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is empathy-induced helping due to self-other merging? *Journal of Personality and Social Psychology*, *73*(3), 495-509. https://doi.org/10.1037/0022-3514.73.3.495

Berant, E., & Wald, Y. (2009). Self-reported attachment patterns and Rorschach-related scores of ego boundary, defensive processes, and thinking disorders. *Journal of Personality Assessment*, *91*(4), 365-372. https://doi.org/10.1080/00223890902936173

Bettis, R. A, Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conservation among authors, referees, and editors. *Strategic Management Journal*, *37*(2), 257-261.

Block, J., & Kuckertz, A. (2018). Seven principles of effective replication studies: strengthening the evidence base of management research. *Management Review Quarterly*, *68*, 355-359. https://doi.org/10.1007/s11301-018-0149-3

Bromet, E., & Moos, R. H. (1977). Environmental Resources and the Posttreatment Functioning of Alcoholic Patients. *Journal of Health and Social Behavior*, *18*(3), 326-338. https://doi.org/10.2307/2136358

Brundidge, J., Baek, K., Johnson, T. J., & Williams, L. (2013). Does the medium still matter? The influence of gender and political connectedness on contacting U.S. public officials online and offline. *Sex Roles*, *69*(1–2), 3-15. https://doi.org/10.1007/s11199-013-0280-5

Buranyi, S. (2017, June 27). Is the staggeringly profitable business of scientific publishing bad for science? Retrieved January 14, 2021, from https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

Burtăverde, V., De Raad, B., & Zanfirescu, A.-Ş. (2018). An emic-etic approach to personality assessment in predicting social adaptation, risky social behaviors, status striving and social affirmation. *Journal of Research in Personality*, *76*, 113-123. https://doi.org/10.1016/j.jrp.2018.08.003

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why a small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cameron, J. E. (2001). Social identity, modern sexism, and perceptions of personal and group discrimination by women and men. *Sex Roles*, *45*, 743-766. https://doi.org/10.1023/A:1015636318953

Chambers, C. (2017). *The seven deadly sins of psychology*. Princeton, NJ: Princeton University Press.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. https://doi.org/10.17605/osf.io/c8akj

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93-99. https://doi.org/10.1016/j.jesp.2015.10.002

Creswell, J. & Clark, P. V. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, Canada: SAGE Publications.

Cutting, J. E. (2007). On the growth of psychological science. *APS observer*, *20*(8), 17-19. https://www.researchgate.net/publication/241265574

Dandeneau, S. D., & Baldwin, M. W. (2004). The inhibition of socially rejecting information among people with high versus low self-esteem: The role of attentional bias and the effects of bias reduction training. *Journal of Social and Clinical Psychology*, *23*(4), 584-602. https://doi.org/10.1521/jscp.23.4.584.40306

Dunlap, K. (1926). The experimental methods of psychology. In: C. Murchison (Ed.) *Psychologies of 1925* (pp. 331-353). Worchester: Clark University Press.

Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' reputations are based on getting it right, not being right. *PLOS Biology*, *14*(5), e1002460. https://doi.org/10.1371/journal.pbio.1002460

Galanis, C. M. B., & Jones, E. E. (1986). When stigma confronts stigma: Some conditions enhancing a victim's tolerance of other victims. *Personality and Social Psychology Bulletin, 12*(2), 169-177. https://doi.org/10.1177/0146167286122003

Galetzka, C. (2019). A short introduction to the reproducibility debate in psychology. *Journal of European Psychology Students*, *10*(3), 16-25. https://doi.org/10.5334/jeps.469

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis – a 'garden of forking paths' – explains why many statistically significant comparisons don't hold up. *American Scientists*, *102*(6), 460.

Griese, E. R., McMahon, T. R., & Kenyon, D. Y. B. (2017). A research experience for American Indian undergraduates: Utilizing an actor–partner interdependence model to examine the student–mentor dyad. *Journal of Diversity in Higher Education, 10*(1), 39-51. https://doi.org/10.1037/a0040033

Gyurcsik, N. C., Brawley, L. R., & Langhout, N. (2002). Acute thoughts, exercise consistency, and coping self-efficacy. *Journal of Applied Social Psychology, 32*(10), 2134-2153. https://doi.org/10.1111/j.1559-1816.2002.tb02067.x

Hevey, D., Pertl, M., Thomas, K., Maher, L., Craig, A., & Ni Chuinneagain, S. (2010). Consideration of future consequences scale: Confirmatory Factor Analysis. *Personality and Individual Differences, 48*(5), 654-657. https://doi.org/10.1016/j.paid.2010.01.006

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology, 97*(6), 963-976. https://doi.org/10.1037/a0017423

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645-654. https://doi.org/10.1177/1745691612464056

Isager, P. M. (2018, June 11). What to replicate? Justifications of study choice from 85 replication studies. *Zenodo*. https://doi.org/10.5281/zenodo.1286715

Isager, P. M. (2019, April 12). Quantifying the corroboration of a finding. Retrieved from https://pedermisager.netlify.app/post/quantifying-the-corroboration-of-a-finding/

Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van 't Veer, A. E., Vranka, M., & Lakens, D. (in press). Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection. https://doi.org/10.31222/osf.io/2gurz

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

Jonas, E., Schimel, J., Greenberg, J., & Pyszczynski, T. (2002). The Scrooge effect: Evidence that mortality salience increases prosocial attitudes and behavior. *Personality and Social Psychology Bulletin, 28*(10), 1342-1353. https://doi.org/10.1177/014616702236834

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *Journal of Psychiatry & Neuroscience, 37*(3), 149-152. https://doi.org/10.1503/jpn.120065

Kalibatseva, Z., & Leong, F. T. L. (2018). Cultural factors, depressive and somatic symptoms among Chinese American and European American college students. *Journal of Cross-Cultural Psychology, 49*(10), 1556-1572. https://doi.org/10.1177/0022022118803181

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Lodge, M., & Taber, C. S. (2005). The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology, 26*(3), 455-482. https://doi.org/10.1111/j.1467-9221.2005.00426.x

Makel, M. C., & Plucker, J. A. (2014). Creativity is more than novelty: Reconsidering replication as a creativity act. *Psychology of Aesthetics, Creativity, and the Arts, 8*(1), 27–29. https://doi.org/10.1037/a0035811

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542. https://doi.org/10.1177/1745691612460688

Mazur, A., Booth, A., & Dabbs, Jr., J. M. (1992). Testosterone and Chess Competition. *Social Psychology Quarterly, 55*(1), 70-77. https://doi.org/10.2307/2786687

Milkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience, 45*(3), 163-172. https://doi.org/10.1007/s10827-018-0702-z

Mischel, W., & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology, 16*(2), 329-337. https://doi.org/10.1037/h0029815

Morrison, K. R. (2011). A license to speak up: Outgroup minorities and opinion expression. *Journal of Experimental Social Psychology, 47*(4), 756-766. https://doi.org/10.1016/j.jesp.2011.03.004

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science (American Association for the Advancement of Science)*, *349*(6251), Aac4716. https://doi.org/10.1126/science.aac4716

Oyamot Jr., C. M., Jackson, M. S., Fisher, E. L., Deason, G., & Borgida, E. (2017). Social norms and egalitarian values mitigate authoritarian intolerance toward sexual minorities. *Political Psychology*, *38*(5), 777-794. https://doi.org/10.1111/pops.12360

P-curve app 4.06 (2017). Retrieved from https://p-curve.com/app4

P-curve results app 4.06 (2017). Retrieved from https://p-curve.com/app4 after clicking on the button "Make the *p*-curve"

Peacock, M. J., Cowan, G., Bommersbach, M., Smith, S. Y., & Stahly, G. (1997). Pretrial predictors of judgments in the O. J. Simpson case. *Journal of Social Issues, 53*(3), 441-454. https://doi.org/10.1111/j.1540-4560.1997.tb02121.x

Puddifoot, J. E. (1996). The persuasive effects of a real and complex communication. *The Journal of Social Psychology*, *136*(4), 447-459. https://doi.org/10.1080/00224545.1996.9714026

R Core Team (2013). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing [Software]. Vienna, Austria. https://www.R-project.org/

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Replication, replication, replication. *Psychologist*, *25*, 346–348.

Rozin, P., Lowery, L., & Ebert, R. (1994). Varieties of disgust faces and the structure of disgust. *Journal of Personality and Social Psychology*, *66*(5), 870-881. https://doi.org/10.1037/0022-3514.66.5.870

Santens, E., Claes, L., Dierckx, E., Luyckx, K., Peuskens, H., & Dom, G. (2018). Personality profiles in substance use disorders: Do they differ in clinical symptomatology, personality disorders and coping? *Personality and Individual Differences, 131*, 61-66. https://doi.org/10.1016/j.paid.2018.04.018

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*(2), 90-100. https://doi.org/10.1037/a0015108

Schwab, A., & Starbuck, W. (2017). A call for openness in research reporting: How to turn covert practices into helpful tools. *Academy of Management Learning & Education*, *16*, 125-141.

Silva, P., Delerue Matos, A., & Martinez-Pecino, R. (2018). Confidant network and quality of life of individuals aged 50+: the positive role of internet use. *Cyberpsychology, Behavior, and Social Networking*, *21*(11), 694-702. https://doi.org/10.1089/cyber.2018.0170

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *P*-curves: Making *P*-curve analysis more robust to errors, fraud, and ambitious *P*-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146-1152. https://doi.org/10.1037/xge0000104

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689-699. https://doi.org/10.1037/0022-3514.52.4.689

Spellman, B. A. (2012). Introduction to the special section: Data, data everywhere … especially in my file drawer. *Perspectives on Psychological Science*, *7*(1), 58-59. https://doi.org/10.1177/1745691611432124

Stellar, J. E., Cohen, A., Oveis, C., & Keltner, D. (2015). Affective and physiological responses to the suffering of others: Compassion and vagal activity. *Journal of Personality and Social Psychology*, *108*(4), 572-585. https://doi.org/10.1037/pspi0000010

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, *12*(1), 53-74.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*(3), 437-446. https://doi.org/10.1037/0022-3514.73.3.437

Strathern, M. (1997). 'Improving ratings': audit in the British University System. *European Review*, *5*(3), 305-321.

Surmann, A. T. (1997). The effects of race, weight, and gender on evaluations of writing competence. *The Journal of Social Psychology*, *137*(2), 173-180. https://doi.org/10.1080/00224549709595428

Thieme, R. E., & Feij, J. A. (1985). Tyramine, a new clue to disinhibition and sensation seeking? *Personality and Individual Differences*, *7*(3), 349-354. https://doi.org/10.1016/0191-8869(86)90010-3

Thomas, K. J., & Mucherah, W. M. (2018). Brazilian adolescents' just world beliefs and its relationships with school fairness, student conduct, and legal authorities. *Social Justice Research*, *31*(1), 41-60. https://doi.org/10.1007/s11211-017-0301-6

Veling, H., Holland, R. W., & Van Knippenberg, A. (2008). When approach motivation and behavioral inhibition collide: Behavior regulation through stimulus devaluation. *Journal of Experimental Social Psychology*, *44*(4), 1013-1019. https://doi.org/10.1016/j.jesp.2008.03.004

Verma, J. P., & Verma, P. (2020). *Determining sample size and power in research studies: A manual for researchers* (1st ed.). Singapore: Springer Publishing. https://doi.org/10.1007/978-981-15-5204-5

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. *Perspectives on Psychological Science*, *11*(6), 917-928. https://doi.org/10.1177/1745691616674458

Waltman, L., & Noyons, E. (2018, March). *Bibliometrics for Research Management and Research Evaluation*. CWTS BV. Retrieved from https://www.cwts.nl/pdf/CWTS_bibliometrics.pdf

Web of Science Core Collection Help. (2020). Retrieved January 01, 2021, from https://images.webofknowledge.com/images/help/WOS/hs_wos_fieldtags.html

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.01832

Zhang, L., Qiu, Y., & Teng, E. (2017). Cross-level relationships between justice climate and organizational citizenship behavior: Perceived organizational support as mediator. *Social Behavior and Personality: An International Journal*, *45*(3), 387-397. https://doi.org/10.2224/sbp.4842

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, 1-50. https://doi.org/10.1017/S0140525X17001972

Appendix A. Overview QRPs and RDF

Table A1 contains an overview of common QRPs and RDF.

Table A1

Overview QRPs and RDF

Description QRP and/or RDF	*Useful references*
Failing to report all dependent measures	John et al., 2012; Simmons et al., 2011
Collecting more data after seeing whether results were significant; Optional stopping of collecting data	John et al., 2012; Wicherts et al., 2016; Simmons et al., 2011
Selectively reporting studies that 'worked'	John et al., 2012
Not reporting all conditions	John et al., 2012; Simmons et al., 2011
Incorrectly rounding off *p* values	John et al., 2012
Selectively excluding data after looking what the results are after exclusion	John et al., 2012; Simmons et al., 2011
Incorrectly claiming a lack of effect of demographic variables	John et al., 2012
Falsification of data; Correcting, coding, and/or discarding data during the collection of data	John et al., 2012; Wicherts et al., 2016
Selecting the dependent variable out of several alternative measures of the same construct	Wicherts et al., 2016
Selecting another construct as the primary outcome	Wicherts et al., 2016
Selecting independent variables out of a set of manipulated independent variables	Wicherts et al., 2016
Including additionally measured variables as covariates, independent variables, mediators, and/or moderators	Wicherts et al., 2016; Simmons et al., 2011
Operationalizing independent variables in different ways (e.g., by discarding or combining levels of factors).	Wicherts et al., 2016
Making ad hoc decisions about dealing with: missing data, outliers, violations of statistical assumptions, in-/exclusion criteria, and pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction)	Wicherts et al., 2016
Making choices about: statistical models, estimation methods, software packages, computation of standard errors, and inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)	Wicherts et al., 2016
Hypothesizing After Results Are Known (HARKing); Presenting post hoc hypotheses as if they were a priori; Presenting exploratory analyses as if they were confirmatory	Kerr, 1998; John et al., 2012; Wicherts et al., 2016

Appendix B. Scoring the Top 10 Studies on the RDF Checklist

In order to map the DFS of the top 10 studies, each of the ten studies is scored on RDF. This Appendix contains the scores for each study on each of the eight items on the RDF checklist.

Number 1 of the Top 10

The number 1 of the top 10 is the first study reported in the paper *Testosterone and Chess Competition* (Mazur et al., 1992). Table B1 shows the score on each of the eight selected RDF for the number 1 paper.

Table B1

Coding paper nr. 1 of the top 10 (i.e., Mazur et al., 1992)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Two confirmatory hypotheses: "We hypothesize that male chess competitors show the same T pattern as seen in physically taxing sports. Specifically, 1) chess players' T will rise just before competition, and 2) afterward the winners' T will be higher than that of the losers." (Mazur et al., 1992, p. 71) "Confirmatory evidence, at first limited to physical athletic competition, now has been extended to nonphysical face-to-face competition, which is only one step removed from normal conversational interaction." (Mazur et al., 1992, p. 76)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"We used repeated-measures ANOVAs to test whether the T patterns of winners and of losers were significantly different. Each subject underwent 24 T measurements: three times per week (the day before each match, just before the match, and just after) over eight weeks (excluding Week 8). Missing T values for winners (losers) are replaced by the mean value of other winners' (losers') values at the corresponding time." (Mazur et al., 1992, p. 74) "An ANOVA treating T1-T8 as repeated measures was used to test whether winners and losers differed on their overall T patterns. Because the ANOVA program (Freund, Littell, and Spector 1986) eliminates subjects with missing data, we replaced missing values for this procedure only. (TI is missing for one winner and is replaced by the mean TI of the other winners. One loser is missing T6,

		another T7, and two T8; these values are replaced by the corresponding means of other losers.)" (Mazur et al., 1992, p. 73)
Sample size (predetermined or not).	2	Regional tournament: "We recruited subjects from among the members of a city chess club at a meeting where we explained the purpose and methods of the research. Subjects gave their informed consent to participate without pay, and they do not differ in any apparent way from nonparticipating club members. We studied 16 male players as they competed along with nonsubject players in one or both of two chess tournaments. Their T was measured from saliva samples taken the day before, just before, and just after each round of each tournament. The two tournaments differed in time duration and in importance to players. The more important and more prestigious was a regional tournament of four rounds, all played in one day, which drew 26 adults (age 18 or older) plus younger players from several states. (Adults and youths do not play against each other.) Nine men ranging in age from 18 to 64 (median age = 33), plus two 16-year-olds, volunteered to be subjects." (Mazur et al., 1992, p. 71) "The 11 subjects in the regional meeting include two 16-year-olds who competed in the youth division. Four subjects, including one 16-year-old, won three or four rounds of the four-round tournament and are counted here as winners; the remaining seven subjects won zero to two rounds and are categorized as losers. The four winners have high four-digit skill ratings, but not all of them are the highest- rated players in our sample: they rank second, third, fourth, and eighth among the 11 subjects." (Mazur et al., 1992, p. 72) City tournament: "Less serious was the annual city tournament, in which the majority of participants were members of the club who knew one another from their play every Thursday night throughout the year. This city tournament consisted of nine rounds, one per week at the usual Thursday meeting, for a total of nine weeks. Of 26 participants, eight volunteered to be subjects (including three who had been subjects in the regional tournament);" (Mazur et al., 1992, p. 71) "Eight subjects who participated in the city tournament ranged in age from 26 to 64. Five

		subjects won more games than they lost (excluding forfeits and byes), and are considered here as winners; three subjects who lost most of their games are counted as losers. All of the winners had higher ratings than the losers." (Mazur et al., 1992, p. 74)
Sharing/Openness (i.e., materials, data, code).	3	Publication year is 1992.
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"Because the rounds of the city tournament are held a week apart, it is reasonable to assume that postgame changes in T due to winning or losing one round would dissipate before they could be perturbed by any prematch rise in anticipation of the next round." (Mazur et al., 1992, p. 74-75)
Fallacious interpretation of (lack of) statistical significance.	3	"Figure 3 shows the T changes of game winners and losers for both close and far games. After close games, the mean T of winners rises, while that of losers falls, as hypothesized. After far games, on the other hand, win or loss has no effect; trends for both outcomes are similar to the trend for losers of close games. (Treating T values just before and just after the game as repeated measures, an ANOVA shows the interaction between win/loss and close/far to be significant; p = .04.)" (Mazur et al., 1992, p. 75)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	The paper does not disclose enough statistics to calculate the single-article *p*-curve. The following statistics cannot be processed by the *p*-curve app ("*P*-curve app 4.06," 2017): "The repeated-measures ANOVA shows a significant interaction between time and winner/loser (p = .002), as expected. Overall, T behaves differently for winners and for losers across the regional tournament. For winners we found a prematch rise in T (from Time 1 to Time 2), as we hypothesized and as is consistent with prior research. A period comparison t-test of T2 minus TI is nearly significant (p = .08, based on only three winners because Ti is missing for the fourth). Also as hypothesized, the winners' T rises above that of losers, significantly so on the morning of the day

after the tournament (Time 7: $p = .03$, t-test). TI, measured in the morning of the day before the tournament, is surprisingly different for the eventual winners and losers. This difference is not only significant ($p = .02$, t-test) but remarkably consistent: All seven losers have higher TI values than any of the winners (we discount one winner for whom TI is missing). This is not an artifact of the normalization procedure because all of the losers' raw TI scores (ranging from 11.0 to 12.4 ng/dl) also are higher than those of winners (ranging from 2.0 to 10.7 ng/dl). It is not clear which group, if either, departs from normal morning values. Nonetheless, in relative terms the eventual losers of the tournament had reliably high T the day before competing. One consequence is that losers' mean T dropped significantly ($p = .01$, paired comparison t-test) from Time 1 to Time 2, the morning of the tournament, a finding that is not consistent with our hypothesis of a prematch rise." (Mazur et al., 1992, p. 73-74)

"A simple 2 x 24 model, comparing winners and losers across these 24 repeated measurements, shows a highly significant interaction between order of measurement and winner/loser ($p = .0001$). A more complex 2 x 3 x 8 ANOVA, using winner/loser by three measurement times each week by eight weeks, shows a significant interaction between week and winner/loser ($p = .02$). Thus both models show that winners and losers have significantly different T patterns. Applying a simple t-test to each point in time shows the difference between winners and losers to be significant after the sixth game ($p = .05$), after the seventh game ($p = .02$), and after the final game ($p = .001$). Overall, T is higher for winners than for losers after the first two weeks of the city tournament." (Mazur et al., 1992, p. 74)

"Figure 3 shows the T changes of game winners and losers for both close and far games. After close games, the mean T of winners rises, while that of losers falls, as hypothesized. After far games, on the other hand, win or loss has no effect; trends for both outcomes are similar to the trend for losers of close games. (Treating T values just before and just after the game as repeated measures, an ANOVA shows the interaction between win/loss and close/far to be significant; $p = .04$.) There is no sign of a pregame rise in T in either close or far games." (Mazur et al., 1992, p. 75)

Number 2 of the Top 10

The number 2 of the top 10 is the third study reported in the paper *The Generality of the Automatic Attitude Activation Effect* (Bargh et al., 1992). Table B2 shows the score on each of the eight selected RDF for the number 2 paper.

Table B2

Coding paper nr. 2 of the top 10 (i.e., Bargh et al., 1992)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	3	"At the least, this memory word aspect of the test of attitude automaticity must be examined before drawing the more general conclusion that the mere observation of an object can automatically elicit an evaluative response." (Bargh et al., 1992, p. 902) "To provide an exact replication of the Fazio et al. (1986) Experiment 2 (300-ms SOA condition), we dropped the consistent prime stimuli used in our own Experiments 1 and 2. Thus, only fast and slow (good and bad) attitude objects served as primes. In all other respects, the materials and apparatus were the same as in those experiments." (Bargh et al., 1992, p. 902) "The results of Experiment 3 indicate that the automatic attitude activation effect is not dependent on the memory word instruction feature of the original paradigm. Thus, when the requirement for subjects to hold the attitude object prime in working memory during each trial of the automaticity test is removed, providing a better test of the "mere presence" hypothesis, the automatic activation effect is still obtained. Moreover, the effect is obtained for the subject's slowest (weakest) as well as his or her fastest (strongest) attitudes. Thus, Experiment 3 is another demonstration that under conditions more closely approximating the mere presence of the object in the environment, the effect occurs for a majority of objects for which one has a stored evaluation. Finally, the reliable three-way interaction, in which the effect was found to be statistically stronger for fast than for slow attitude object primes, replicated the findings of Experiment 1, in which the attitude assessment phase also came immediately before the test of automaticity." (Bargh et al., 1992, p. 903)

Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	"As part of a mass testing session at the beginning of the semester, 274 introductory psychology students at New York University (NYU) completed a Semantic Evaluations Questionnaire (SEQ)." (Bargh et al., 1992, p. 895) "Fifty-nine NYU introductory psychology students participated in Experiment 3 as partial fulfillment of a course requirement." (Bargh et al., 1992, p. 902)
Sharing/Openness (i.e., materials, data, code).	2	Data are shared in the Appendix (Bargh et al., 1992, p. 912).
Using covariates and reporting the results with and without the covariates.	0	"We began this line of research by collecting normative data on characteristics of attitude object stimuli that might covary with speed of object evaluation or associative strength. As shown in Table 1, all of these factors correlated reliably with mean evaluation latency. However, a correlation with mean attitude object evaluation latency is not the same as a correlation with the automatic activation effect itself. Therefore, using the data from Experiments 1-3, we examined the extent to which these factors moderated the automaticity effect; that is, the signature Prime Valence x Target Valence interaction" (Bargh et al., 1992, p. 903)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	
Fallacious interpretation of (lack of) statistical significance.	1	"Experiment 3 was also a replication of the basic paradigm, this time with the alteration of removing the memory word instructions for half of the subjects. By having subjects hold the attitude object prime in memory until they had evaluated the

adjective target on each trial in the priming task, the original paradigm did not permit unambiguous conclusions about whether the automatic activation effect would occur without such deliberate, intensive conscious thought about the attitude object. Of course, if the effect required only the mere presence of the attitude object to occur, such memory word instructions should not be necessary to produce the effect in the laboratory. The results of Experiment 3 showed that indeed they were not necessary; the automatic activation effect held for both the subject's fastest and slowest evaluated attitude objects." (Bargh et al., 1992, p. 907)

"We conducted further statistical tests of whether the automaticity effect (i.e., the simple Prime Valence x Target Valence interaction) held reliably across the three experiments for the subjects' slow and fast attitude object primes as well as for the consistent but slow primes used in Experiments 1 and 2.13 To do so we used meta-analytic procedures for combining results across studies and for assessing differences in the size of the effect across studies (Rosenthal, 1978; Rosenthal & Rubin, 1979). Not surprisingly, given its reliability in each of the three studies individually, the automaticity effect for the fast attitude objects was reliable overall, with weighted average $Z = 4.42$, $p < .001$, average effect size = .43; moreover, the effect sizes did not differ across the three experiments, $\chi^2(2, N = 106) = 0.29$, $p = .86$. Consistent attitude objects presented to subjects in Experiments 1 and 2 (see Table 2) also showed a reliable automaticity effect, $Z = 4.11$, $p < .001$, average effect size = .60, which did not vary across the two studies, $\chi^2(1, N = 47) = .06$, $p = .81$. Of greatest importance, however, the same pattern of results was obtained in the case of the slow attitude object primes. The automaticity effect was reliable, $Z = 2.66$, $p < .01$, effect size = .26, and it did not vary reliably across the three experiments, $\chi^2(2, N = 106) = 0.52$, $p = .77$. Finally, comparisons between prime types showed that the automaticity effect was reliably greater for fast than slow primes ($p < .004$) and also reliably greater for consistent than slow primes ($p < .03$). However, the automaticity effect proved nonreliably smaller for fast than consistent primes ($Z < 1$), even though the latter primes were associated with reliably slower response times than the fast primes. Taken together, the results of the three experiments suggest the automatic attitude

		activation effect is quite general, holding across most if not all of the range of 92 attitude objects that served as stimuli. These attitude objects varied widely as to their extremity, ambivalence, and polarization of attitude, their consistency of evaluation across subjects (i.e., consensus), and their mean evaluation latencies (see Appendix). Moreover, removing features of the paradigm that potentially could have contributed to the automaticity effect did not change its strength across the three experiments, demonstrating its relatively unconditional nature (see Bargh, 1989)." (Bargh et al., 1992, p. 907)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"Not surprisingly, subjects instructed to remember the prime word on each trial recalled reliably more primes (M = 10.4, 65%) than did subjects not given memory instructions (M = 9.1, 57%); $t(57) = 2.09$, p = .04." (Bargh et al., 1992, p. 903) "As in the previous experiments, the mean facilitation scores for each of the 14 experimental conditions were computed and entered into a 2 (memory word instructions: yes vs. no) x 2 (prime type) x 2 (prime valence) x 2 (target valence) ANOVA. The Prime Valence x Target Valence interaction was again highly reliable, $F(l, 57) = 33.72$, $p < .001$, $MS_e = 5{,}964.7$, accounting for 4.4% of the total variance." (Bargh et al., 1992, p. 903) "This two-way interaction was qualified, however, by a reliable three-way interaction involving prime type, $F(l, 57) = 12.37$, $p = .001$, $MS_e = 4{,}001.0$ (1.1% of total variance). Within this three-way interaction, simple effects tests revealed that the simple Prime Valence x Target Valence interaction was reliable for fast primes, $F(l, 57) = 56.98$, $p < .001$, and also for slow primes, $F(l, 57) = 6.32$, $p = .02$" (Bargh et al., 1992, p. 903) "There was also a main effect for target valence, $F(l, 57) = 24.74$, $p < .001$ (9.6% of variance), which was qualified by a Target Valence x Prime Type interaction, $F(l, 57) = 11.57$, $p = .002$ (0.7% of variance). As can be seen in Figure 3, the general tendency for negative targets to have faster evaluation times than positive targets was stronger in the slow prime condition. All other main effects and interactions were nonsignificant at $p > .20$." (Bargh et al., 1992, p. 903) "Normative polarization was marginally related to the effect ($p = .08$)." (Bargh et al., 1992, p. 905) "In the simultaneous analysis, normative evaluation latency was reliable ($t = 2.20$, $p < .03$), and

ambivalence proved marginally reliable (t = 1.65, p < .10).” (Bargh et al., 1992, p. 905)
“This predictor accounted for nearly all of the explained variance in the regression, t = 71.71, p <. 0001, R^2 = . 36.” (Bargh et al., 1992, p. 905)
“The correlations among the normative predictors can be found in Table 1. Idiosyncratic prime evaluation latency also correlated moderately but reliably (all ps < .001) with the normative predictors across all trials in the regression data set (rs = - .36 with consensus, .47 with normative latency, -.34 with extremity, .12 with ambivalence, and -.23 with polarization).”(Bargh et al., 1992, p. 905)
“With individual entry at Step 8, both normative prime evaluation latency (t = 2.24, p < .025) and idiosyncratic prime evaluation latency (t = 2.13, p < .04) were reliable (all other ps > .25). As noted earlier, however, significance of a predictor when entered by itself might be caused by its relation to another predictor. When the independent influences of normative and idiosyncratic evaluation latencies were assessed in the simultaneous-entry version of the analysis, normative evaluation latency remained (marginally) reliable (t = 1.83, p < .07), whereas idiosyncratic evaluation latency did not (t = 1.30, p = .20), confirming the outcome of the original regression analysis.” (Bargh et al., 1992, p. 906)

In Figure B1, the results are shown of entering the following statistics into the online *p*-curve app (“*P*-curve app 4.06,” 2017): $t(57)$ = 2.09; $F(1,57)$ = 33.72; $F(1,57)$ = 12.37.

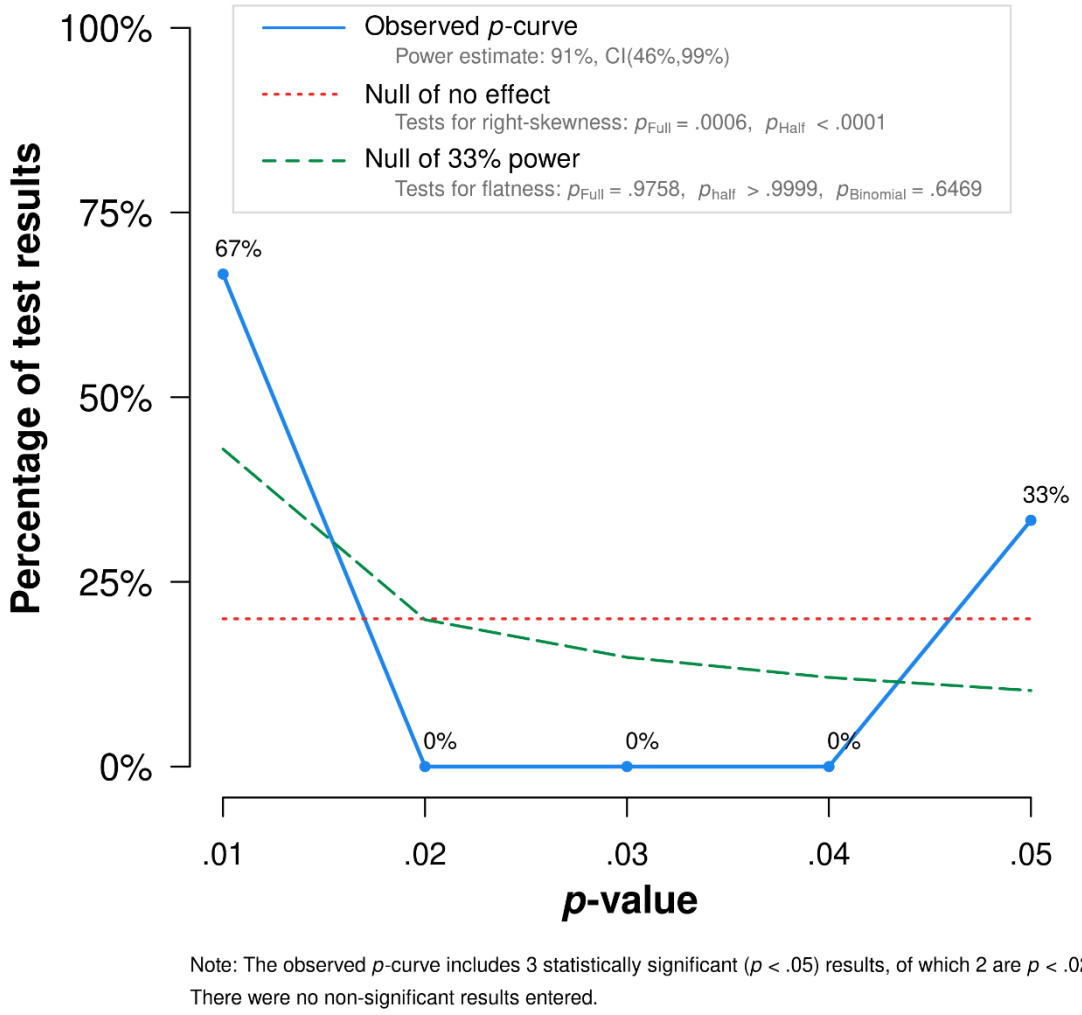


Figure B1. The single-article *p*-curve for the number 2 of the top 10 (i.e., Bargh et al., 1992).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B2, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p = .0006$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B2, the 33% power test is $p = .9758$ for the full *p*-curve, for the half p-curve is $p = .9999$, and for the binomial 33%

power test is $p = .6469$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.5	Z=-3.22, p=.0006	Z=-4.27, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p=.6469	Z=1.97, p=.9758	Z=3.78, p=.9999
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 91% 90% Confidence interval: (46% , 99%)		

Figure B2. Additional statistics for the single-article *p*-curve for the number 2 of the top 10 (i.e., Bargh et al., 1992).

Number 3 of the Top 10

The number 3 of the top 10 is the first study reported in the paper *The Scrooge Effect: Evidence That Mortality Salience Increases Prosocial Attitudes and Behavior* (Jonas et al., 2002). Table B3 shows the score on each of the eight selected RDF for the number 3 paper.

Table B3

Coding paper nr. 3 of the top 10 (i.e., Jonas et al., 2002)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "From a TMT perspective, then, prosocial behavior that conforms to one's personalized belief system would offer the best protection against existential fear. Based on this reasoning, Study 1 was designed to test the hypothesis that a subtle real-world reminder of mortality would increase the favorability of people's attitudes toward charities." (Jonas et al., 2002, p. 1344) "Following Pyszczynski et al. (1996), Study 1 was a field study in which pedestrians walking on the street were interviewed in front of a funeral home or several blocks away from the funeral home. Their attitudes toward two different charitable organizations that they deemed moderately important constituted the dependent variable. We predicted that mortality salient participants, that is, those interviewed in front of the funeral home, would exhibit more favorable attitudes toward the

		two charities than would participants interviewed away from the funeral home." (Jonas et al., 2002, p. 1345) "Of interest, whereas most prior TMT research has focused on negative or socially destructive consequences of confronting one's mortality, such as prejudice, bias, and aggression, Dickens's story hypothesizes a constructive consequence of mortality salience: If generous behavior helps to restore the belief that one is a meaningful and valuable contributor to one's cultural conception of reality, then reminders of mortality should encourage people (perhaps even the "Scrooges" of the world) to be kinder and more benevolent to others. The primary purpose of the two studies reported here was to assess this Dickensian hypothesis." (Jonas et al., 2002, p. 1343) The second footnote: "A second hypothesis also was explored in this first study. We were initially interested in whether mortality salience would affect preferences for the chosen versus nonchosen charity. This hypothesis was derived from cognitive dissonance research showing that after people make an important decision between two closely valued alternatives, the chosen alternative will be highly favored over the nonchosen alternative. There was no support for this spreading of choice alternatives, either with or without mortality salience, thus, the classical free choice dissonance effect was not replicated, which rendered our attempted test of the terror management variation on it ambiguous. Therefore, this unsupported hypothesis is not discussed further." (Jonas et al., 2002, p. 1351)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	"The participants were 17 male and 14 female pedestrians who were solicited to take part in a short survey while walking down a street in

		Boulder, Colorado. All of the participants were U.S. citizens." (Jonas et al., 2002, p. 1345)
Sharing/Openness (i.e., materials, data, code).	2	The first footnote: "The 10 charities used in Study 1 were as follows: Partnership for a Drug-Free America, Association for Retarded Citizens, Acid Rain Foundation, National Alliance to End Homelessness, American Society for the Prevention of Cruelty to Animals, Congress on Racial Equality, Albert Einstein Peace Prize Foundation, Keep America Beautiful, American Foundation for Aging Research (AFAR), and the National Child Safety Council." (Jonas et al., 2002, p. 1351)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	
Fallacious interpretation of (lack of) statistical significance.	3	"From the perspective of terror management theory, reminders of mortality should intensify the desire to express culturally prescribed prosocial attitudes and engage in culturally prescribed prosocial behaviors. Two studies supported these hypotheses. In Study 1, people were interviewed in close proximity to a funeral home or several blocks away and were asked to indicate their attitudes toward two charities they deemed important. Those who were interviewed in front of the funeral home reported more favorability toward these charities than those who were interviewed several blocks away." (Jonas et al., 2002, p. 1342) "The findings of Study 1 extend previous findings that mortality salience increases rewards recommended for a hero and punishment recommended for a moral transgressor (Florian & Mikulincer, 1997; Ochsmann & Reichelt, 1994; Rosenblatt et al., 1989) by showing a similar effect of mortality salience on the value people place on charitable organizations." (Jonas et al., 2002, p. 1346) "Study 1 demonstrated that reminding people of death leads to a more favorable attitude toward

		charities." (Jonas et al., 2002, p. 1349) "Whereas Study 1 showed the effect of mortality salience on attitudes toward charities in the context of a field study in which people were interviewed either in front of or several blocks away from a funeral home," (Jonas et al., 2002, p. 1349) "This research provides evidence of mortality salience affecting yet another type of human behavior: prosocial action. In doing so, the present findings also add to a small but growing body of evidence of behavioral effects of mortality salience" (Jonas et al., 2002, p. 1349) "We believe that by demonstrating positive effects of mortality salience, this work provides an initial step toward an important new direction for terror management research." (Jonas et al., 2002, p. 1349)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	3	"A *t* test performed on this favorability composite yielded a significant effect of our mortality salience treatment, $t(31) = 2.06$, $p < .05$, indicating that mortality salience increased the favorability of participants' attitudes toward the charitable organizations." (Jonas et al., 2002, p. 1345) The third footnote: "We also performed t tests on the individual item composites. This analysis revealed a significant effect of mortality salience for the first and third item composite (i.e., "How beneficial is this charity to society" and "How desirable is this charity to you personally"), $t(31) = 2.08$, $p < .05$ and $t(31) = 1.99$, $p < .06$. A *t* test performed on the second item (i.e., "How much does society need this charity") revealed the same pattern as the one with the other two items but was not statistically significant, $t(31) = 1.38$, $p < .19$." (Jonas et al., 2002, p. 1351)

In Figure B3, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $t(31) = 2.06$; $t(31) = 2.08$; $t(31) = 1.99$; $t(31) = 1.38$.

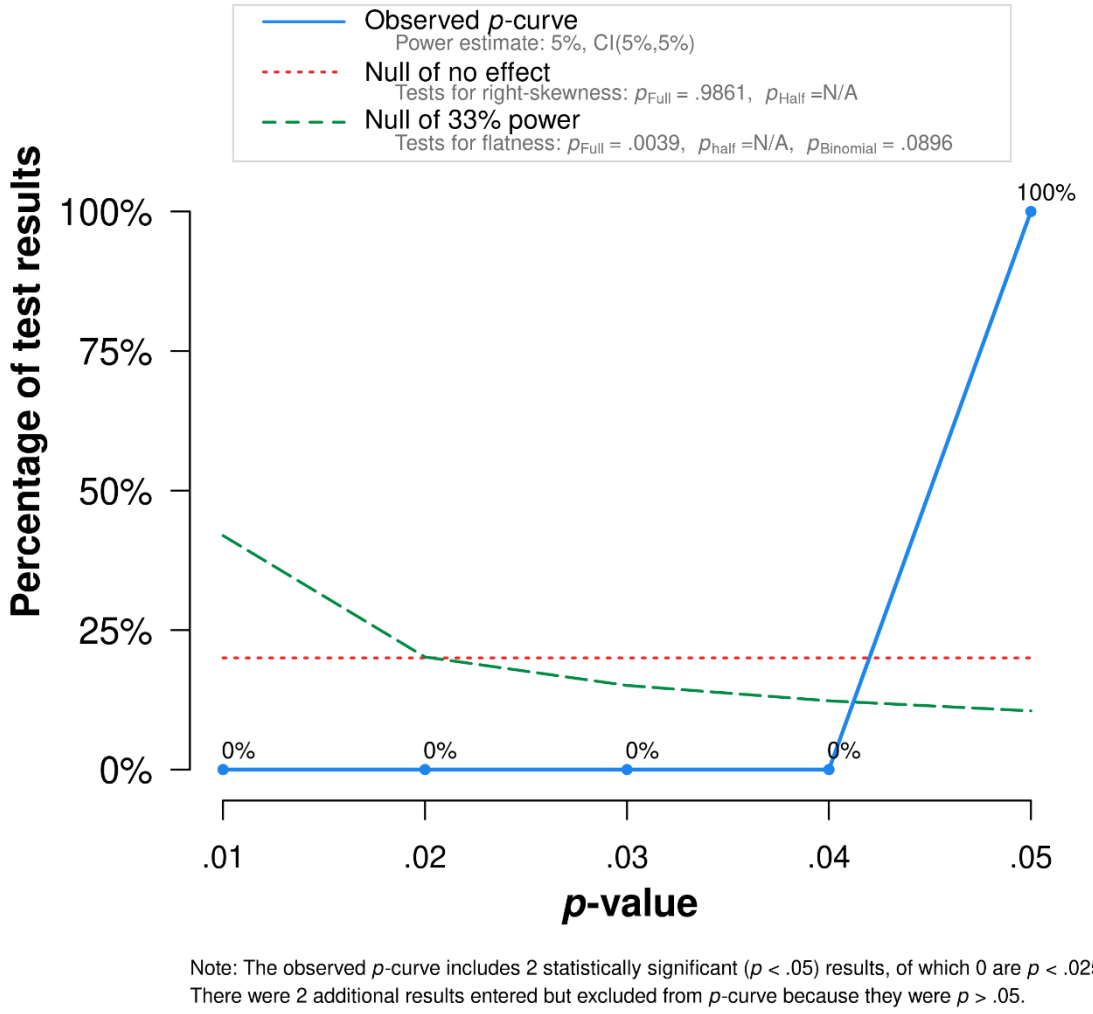


Figure B3. The single-article *p*-curve for the number 3 of the top 10 (i.e., Jonas et al., 2002).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B4, the half *p*-curve test ($p < .0001$) unknown, but the full test ($p = .9861$) is not significantly right-skewed ($p < .1$), which implies that the study does not contain evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B4, the 33% power test is $p = .0039$ for the full *p*-curve, $p = .0896$ for the binomial 33% power test, and unknown for the half *p*-curve. So the *p*-curve indicates that evidential value in this study is inadequate or absent ("*P*-curve results app 4.06," 2017).

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p>.9999	Z=2.2, p=.9861	Z=N/A, p=N/A
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p=.0896	Z=-2.66, p=.0039	Z=N/A, p=N/A (No p<.025 entered)
	Statistical Power		
Power of tests included in p-curve (correcting for selective reporting)	Estimate: 5% 90% Confidence interval: (5% , 5%)		

Figure B4. Additional statistics for the single-article *p*-curve for the number 3 of the top 10 (i.e., Jonas et al., 2002).

Number 4 of the Top 10

The number 4 of the top 10 is the third study reported in the paper *Varieties of Disgust Faces and the Structure of Disgust* (Rozin et al., 1994). Table B4 shows the score on each of the eight selected RDF for the number 4 paper.

Table B4

Coding paper nr. 4 of the top 10 (i.e., Rozin et al., 1994)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	3	"These confounds are inherent in the structure of the face and its musculature. In this study, to remove the confounds, we constructed isolated expressions by cutting out and reassembling segments of two sets of faces (G and M). In this way, we created three faces that were identical (within poser) except for the critical AUs. We then repeated the procedures of Experiment 2 using these two sets of isolated faces." (Rozin et al., 1994, p. 879) "The data generated by the isolated faces show a somewhat more distinct effect than the results from Experiment 2, with a clearer indication of the linkage between upper lip raise and expanded disgust. In general, the results support the three-component analysis. (…) The results from three different studies, involving four different posers and 3-12 face exemplars all indicate that the three principal components of the disgust expression carry different meanings." (Rozin et al., 1994, p. 879)
Exclusion of participants (how	3	Note beneath Table 2: " n = 120." (Rozin et al., 1994, p. 878)

many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.		Unclear why one participant is excluded.
Sample size (predetermined or not).	3	Experiment 3: n = 121. "The subjects for this study were 121 University of Pennsylvania students taking an introductory psychology class. The questionnaire was administered as part of a class project on the recognition of facial expressions of emotion. The results were shared with the class." (Rozin et al., 1994, p. 879)
Sharing/Openness (i.e., materials, data, code).	2	Materials: "Figure 3. Face displays used in Experiment 3. For both posers (M shown here), the face on the left represents an isolated nose wrinkle (Facial Action Coding System Action Unit [AU] 9), the face in the middle an isolated gape (AU26) and tongue extension (AU 19), and the face on the right an isolated upper lip retraction (AU10)." (Rozin et al., 1994, p. 880) Part of data: "Table 2. Subject Response Percentages for Experiments 1, 2, and 3" (Rozin et al., 1994, p. 878)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	
Fallacious interpretation of (lack of) statistical significance.	2	"The study of the development of disgust has been hampered by the lack of a good nonverbal measure. This study may be the first step in providing such a handle. We have demonstrated that Americans take different meanings from different components of the disgust expression. The communicative value of nose wrinkle and gape, which indicate the modality

		being offended, probably derives from the functional value of these responses, as described by Peiper (1963; see introduction to this article).” (Rozin et al., 1994, p. 880) “Upper lip retraction (Face L4 in Figure 1; Face 3 in Figure 3), in all three studies, assumes the burden of communicating the presence of elicitors that would fit under expanded disgust: reminders of animal origins, interpersonal contamination, and moral offense. This effect is extremely clear in Experiments 1 and 3 and present but less compelling in Experiment 2. The linkage of the upper lip raise to anger and contempt in all three studies confirms a link between expanded disgust, including moral disgust, and these two other moral emotions.” (Rozin et al., 1994, p. 880)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	The paper does not disclose enough statistics to calculate the single-article *p*-curve. The following statistics cannot be processed by the *p*-curve app (“*P*-curve app 4.06,” 2017): “The results are shown in the columns on the right of Table 2. There were significant departures from randomness in choices for most of the situations, with most results significant at $p < .001$. Nose wrinkle was the dominant response for both examples of bad smells and bad tastes (sour and bitter). The combination of gape and tongue extension was dominant for "eating a half teaspoon of hot pepper" and was also the (nonsignificantly) most frequent response for "eating an apple with a worm in it." Upper lip raise was the predominant response for disgust, and all of the indicators of expanded disgust body violations and death, interpersonal disgust, and moral violations), with all effects (except the interpersonal effect of "sleeping in hotel bed on which the linens have not been changed") significant at $p < .01$ or better. Upper lip raise was also the dominant response for anger and contempt.” (Rozin et al., 1994, p. 879) “Table 2. Subject Response Percentages for Experiments 1, 2, and 3” (Rozin et al., 1994, p. 878)

Number 5 of the Top 10

The number 5 of the top 10 is the third study reported in the paper *Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility* (Strack & Mussweiler, 1997). Table B5 shows the score on each of the eight selected RDF for the number 5 paper.

Table B5

Coding paper nr. 5 of the top 10 (i.e., Strack & Mussweiler, 1997)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	"The third study tested implications of the accessibility notion for situations in which the comparative judgments are not assumed to involve the activation of relevant information. Specifically, we tested predictions about the different cognitive mechanisms for plausible and implausible anchor values." (Strack & Mussweiler, 1997, p. 439) "Study 3 demonstrated that generating absolute judgments requires more time when comparative judgments include an implausible anchor and can therefore be made without relevant target information that would otherwise be accessible." (Strack & Mussweiler, 1997, p. 437) "The results of Study 3 provide further support for the hypothesis that mechanisms of semantic priming may be responsible for anchoring effects. Specifically, we found that the comparative task took more time when its solution was assumed to require an elaborate test (i.e., for plausible anchors) than when it was assumed to be solvable on the basis of categorical knowledge (i.e., for implausible anchors)." (Strack & Mussweiler, 1997, p. 443)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"Two participants were excluded from the analysis because of missing data." (Strack & Mussweiler, 1997, p. 442)
Sample size (predetermined or not).	3	"We recruited 69 students at the University of Würzburg as participants and asked them to participate in a pretest for the construction of a questionnaire assessing general knowledge. A chocolate bar was offered as compensation." (Strack & Mussweiler, 1997, p. 442) "The questions used were similar to those of Studies 1 and 2. The anchors were chosen such that they differed in both their direction and their

		plausibility. The latter was determined by asking 40 different participants to assess the plausibility of comparative questions by using plausible and implausible anchors on a 5-point rating scale ranging from 1 (absolutely implausible) to 5 (very plausible). Plausible anchors deviated about 1 standard deviation from the mean of the calibration group (n = 151); implausible anchors deviated from this mean by more than 10 standard deviations, except in instances in which such an extreme deviation yielded logical inconsistencies. In addition, for any anchor to qualify as plausible or implausible, more than 80% of the participants had to assign the potential anchor to one of the two extreme categories on the rating scale. As a result, four different types resulted from the orthogonal combination of plausibility (plausible vs. implausible) and direction (high vs. low). They are listed in Table 5." (Strack & Mussweiler, 1997, p. 442) "n = 67 for all cells." (Strack & Mussweiler, 1997, p. 443)
Sharing/Openness (i.e., materials, data, code).	2	Materials in Appendix: "Mean Values for Individual Questions Used in Studies 1 Through 3" (Strack & Mussweiler, 1997, p. 446)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	
Fallacious interpretation of (lack of) statistical significance.	3	Calling an effect substantial based on significance: "Table 6 reveals that a substantial anchoring effect again was found: Overall, high anchors led to higher absolute judgments than did low anchors." (Strack & Mussweiler, 1997, p. 442) "The results of Study 3 also speak to the question of whether anchoring in the implausible condition occurs as a simple adjustment to the boundary of the plausibility range. Perhaps judges simply select the first plausible value of their subjective distribution. The latency data suggest that this

		might not be the case. The fact that finding the absolute answer took more time after comparing the target with an implausible than with a plausible anchor suggests that judges may not have simply selected a boundary value but instead engaged in a more elaborative test. At this point, the data are merely suggestive, and more research about this issue is needed." (Strack & Mussweiler, 1997, p. 443) "Thus both assimilation and contrast effects are possible manifestations of anchoring and must be studied with respect to the underlying cognitive mechanism." (Strack & Mussweiler, 1997, p. 444)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	Study 3: "Table 6 reveals that a substantial anchoring effect again was found: Overall, high anchors led to higher absolute judgments than did low anchors. This difference yielded a significant main effect of direction in a 2 (high vs. low anchors) x 2 (plausible vs. implausible) within-subjects multivariate analysis of variance (MANOVA) with the standardized answers to the eight critical absolute questions as dependent variables, $F(1, 66) = 7.61$, $p < .01$. Moreover, plausibility yielded no significant effect: Anchoring occurred for plausible as well as implausible anchors, $F(1, 66) = 3.55$, $p < .07$, for the interaction and $F(1, 66) < 1$, for the main effect. A mere inspection of the means reveals that implausible anchors were at least as effective as plausible ones." (Strack & Mussweiler, 1997, p. 442-443) "The corresponding two-way Plausibility x Response Type interaction proved to be significant in a 2 (high vs. low anchors) x 2 (plausible vs. implausible anchors) x 2 (comparative vs. absolute responses) MANOVA with transformed response latencies for the answers to the eight critical comparative and absolute questions as dependent variables, $F(1, 66) = 72.81$, $p < .001$. Moreover, plausible anchors yielded shorter response latencies than implausible anchors, $F(1, 66) = 23.78$, $p < .001$, and absolute questions yielded shorter response latencies than comparative questions, $F(1, 66) = 38.81$, $p < .001$. Response latencies did not differ for high and low anchors: $F(1, 66) = 1.35$, $p < .25$, for the main effect of direction; $F(1, 66) = 1.88$, $p < .2$, for the two-way Direction × Plausibility interaction; $F(1, 66) = 2.55$, $p < .12$, for the two-way Direction × Response Type interaction; and $F(1, 66) < 1$, for the three-way interaction. Separate

> analyses for comparative and absolute questions revealed that the main effect of plausibility in a 2 (high vs. low anchors) × 2 (plausible vs. implausible anchors) MANOVA was significant for both response types: For the comparative task, $F(1, 66) = 8.46$, $p < .005$, and for the absolute task, $F(1, 66) = 61.31$, $p < .001$. Again, the response latencies did not differ for high versus low anchors: Neither the corresponding main effects, $F(1, 66) < 1$, for the comparative task and $F(1, 66) = 2.62$, $p < .11$, for the absolute task nor the interaction effects, $F(1, 66) < 1$, for the comparative task and $F(1, 66) = 1.70$, $p < .20$ for the absolute task, attained significance. These results suggest that, for plausible anchors, participants solve the comparative task by engaging in an elaborate and time-consuming test." (Strack & Mussweiler, 1997, p. 443)

In Figure B5, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(1, 66) = 7.61$; $F(1, 66) = 3.55$; $F(1, 66) = 72.81$; $F(1, 66) = 23.78$; $F(1, 66) = 38.81$; $F(1, 66) = 1.88$; $F(1, 66) = 2.55$.

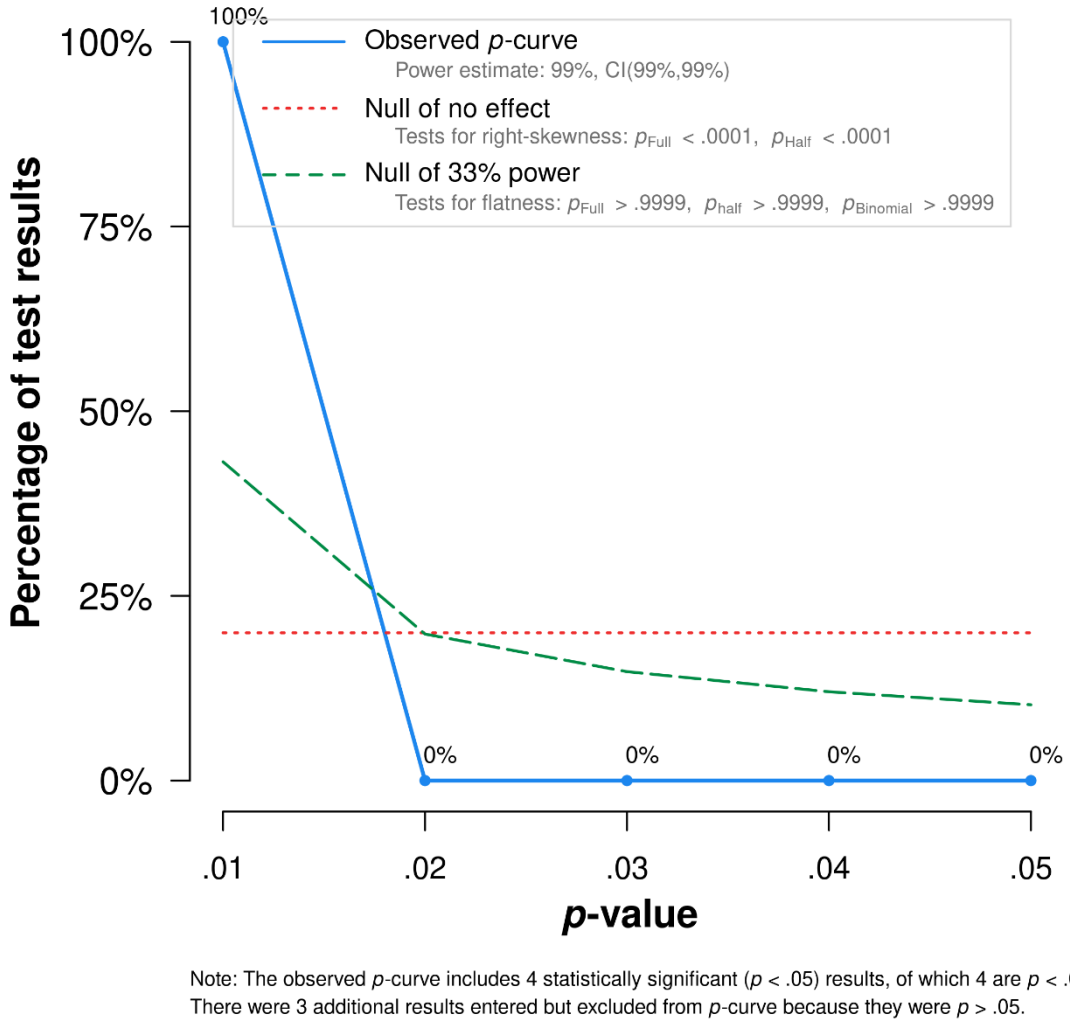


Figure B5. The single-article *p*-curve for the number 5 of the top 10 (i.e., Strack & Mussweiler, 1997).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B6, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($ps < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B6, the 33% power test is $p > .9999$ for the full *p*-curve, for the half *p*-curve is $p > .9999$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.0625	Z=-7.96, p<.0001	Z=-7.49, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p>.9999	Z=6.19, p>.9999	Z=6.77, p>.9999
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure B6. Additional statistics for the single-article *p*-curve for the number 5 of the top 10 (i.e., Strack & Mussweiler, 1997).

Number 6 of the Top 10

The number 6 of the top 10 is the first study reported in the paper *Attention in Delay of Gratification* (Mischel & Ebbesen, 1970). Table B6 shows the score on each of the eight selected RDF for the number 6 paper.

Table B6

Coding paper nr. 6 of the top 10 (i.e., Mischel & Ebbesen, 1970)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "In accord with the previously discussed theoretical ideas, it was predicted that conditions in which the delayed reward was present and visually available would enhance attention to it and hence increase voluntary delay time for it. It was anticipated that the condition in which the child was left without either reward would make it most difficult to bridge the delay time and therefore lead to the shortest waiting. In addition it was expected, although less confidently, that the condition in which both the delayed and immediate reward were available for attention would best facilitate waiting time. This condition might permit the subject to compare and contrast the two outcomes, possibly providing himself with persuasive arguments and self-instructions to help him delay long enough to achieve his preferred gratification. On the other hand, one might also plausibly expect maximum delay when the child could focus his attention on

		the delayed reward without being tempted by the immediate gratification – that is, the condition in which the delayed reward was present for attention but the immediate one was not." (Mischel & Ebbesen, 1970, p. 331-332) "In accord with the previously discussed theorizing, it was expected that as the degree of attention paid to the delayed rewards increased, the length of time which the children waited would increase" (Mischel & Ebbesen, 1970, p. 333)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"The subjects were 16 boys and 16 girls attending the King Nursery School of Stanford University. Three other subjects were run but eliminated because of their failure to comprehend the instructions as described later." (Mischel & Ebbesen, 1970, p. 332) "Three children were unable to answer these questions correctly and were therefore excluded from the data a priori." (Mischel & Ebbesen, 1970, p. 333) "Previous research on preference for delayed rewards has been conducted mainly with subjects at least 6 years of age or older. Preliminary observations of the actual waiting behavior of nursery school children suggested, however, that the capacity to wait for longterm goals and to inhibit both immediate gratification and motoric activity seems to develop markedly at about ages 3-4. It was hoped, therefore, that research with subjects in this young age range should be especially informative in revealing some of the processes that underlie the genesis of goal-directed waiting. A first requirement was a paradigm in which such very young children would be willing to remain in an experimental room, waiting entirely alone for at least a short time without becoming upset and debilitatingly anxious." (Mischel & Ebbesen, 1970, p. 331)
Sample size (predetermined or not).	3	"The subjects were 16 boys and 16 girls attending the King Nursery School of Stanford University. Three other subjects were run but eliminated because of their failure to comprehend the instructions as described later. The children ranged in age from 3 years, 6 months, to 5 years, 8 months (with a median age of 4 years, 6 months). The procedures were conducted by two male experimenters. Eight subjects (4 males and 4 females) were assigned randomly to each of the four experimental conditions. In each condition

		each experimenter ran 2 males and 2 females in order to avoid systematic biasing effects from sex or experimenters." (Mischel & Ebbesen, 1970, p. 332)
Sharing/Openness (i.e., materials, data, code).	2	Table 3: "Number of Children Waiting the Maximum Time (15 minutes) in Each Attention Condition" (Mischel & Ebbesen, 1970, p. 334) Materials: "The experimenter asked the child which of the two rewards he liked better, and after the child chose, said:" (Mischel & Ebbesen, 1970, p. 332)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	
Fallacious interpretation of (lack of) statistical significance.	3	Calling the waiting time much longer based on statistical significance: "Thus, children waited much longer for rewards when the rewards were absent than when any rewards were left available for attention." (Mischel & Ebbesen, 1970, p. 333) "The lack of significant difference in waiting time when the subjects faced the immediate reward or the delayed one does seem understandable from the perspective of frustrative nonreward theory." (Mischel & Ebbesen, 1970, p. 336)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	2	"An analysis of variance of the mean delay times (Table 2) demonstrated that the overall effect of attentional conditions was significant ($F = 4.42$, df = 3/28, $p < .025$). To determine the relative contribution of the conditions to the overall effect, orthogonal contrasts were computed (Winer, 1962). The first orthogonal contrast (Ci in Table 2) compared the effect of having any reward present for attention with having no reward present during the delay period. This comparison yielded an F of 9.52 ($p < .005$, df = 1/28). Thus, children waited much longer for rewards when the rewards were absent than when any rewards were left available for attention. The second orthogonal contrast (C_2)

> compared mean delay times when both rewards were present with mean delay times when either the delayed or the immediate reward was available for attention. The results of this contrast suggested a slight trend toward shorter delay when both rewards were present for attention, rather than when only one reward was present ($F = 3.45$, df = 1/28, $p < .1$). The final contrast, (C_3), comparing attention to the delayed reward with attention to the immediate reward, was not statistically significant ($F < 1$)." (Mischel & Ebbesen, 1970, p. 333-334)
>
> "Table 3 shows the number of subjects in each condition who waited the full 15 minutes. An overall frequency analysis yielded a significant chi-square ($\chi^2 = 11.07$, $p < .025$, df = 3)." (Mischel & Ebbesen, 1970, p. 334)

In Figure B7, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(3, 28) = 4.42$; $F(1, 28) = 9.52$; $F(1, 28) = 3.45$.

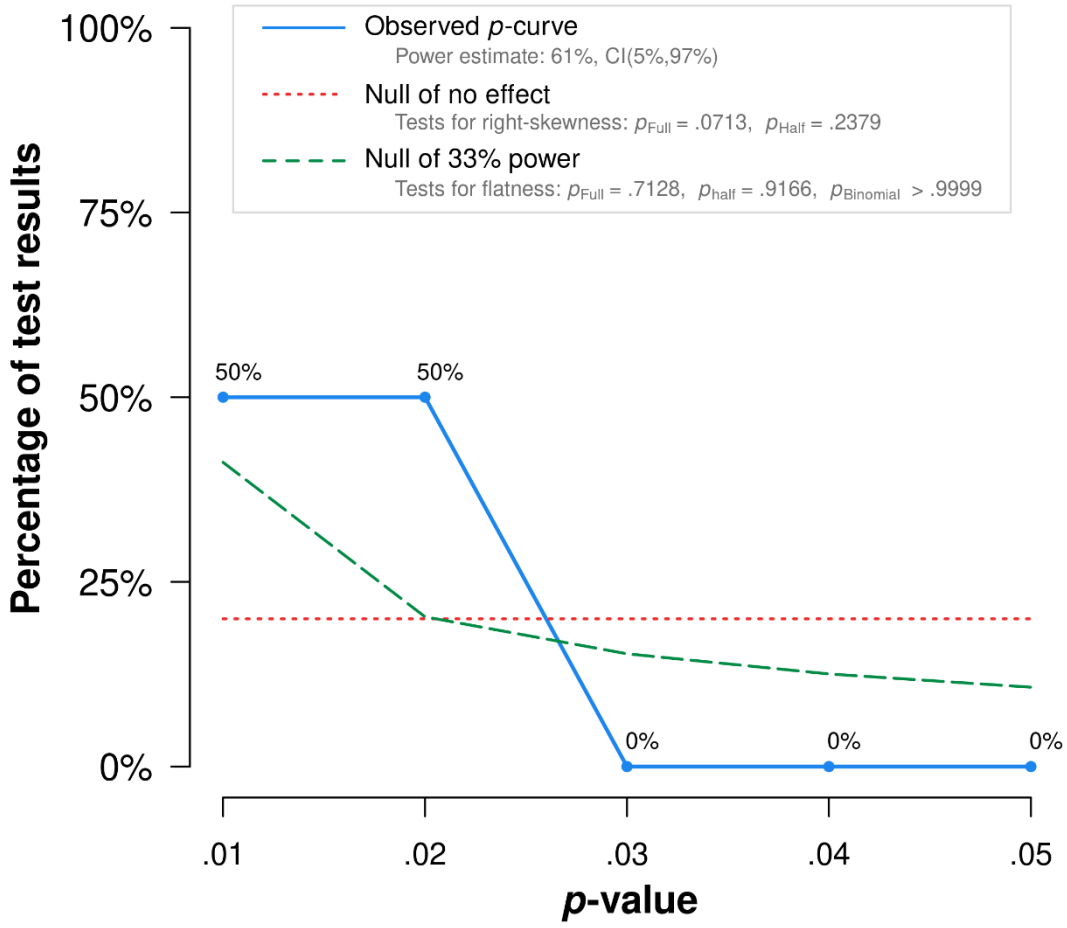


Figure B7. The single-article *p*-curve for the number 6 of the top 10 (i.e., Mischel & Ebbesen, 1970).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B8, not only is the half *p*-curve test ($p = .2379$) not significantly right-skewed ($p < .05$), but also both the half ($p = .2379$) and full test ($p = .0713$) are not significantly right-skewed ($ps < .1$), which implies that the study does not contain evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B8, the 33% power test is $p = .7128$ for the full *p*-curve, for the half *p*-curve is $p = .9166$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	Continuous Test *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. *(Right skew)*	p=.25	Z=-1.47, p=.0713	Z=-0.71, p=.2379
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p>.9999	Z=0.56, p=.7128	Z=1.38, p=.9166
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 61% 90% Confidence interval: (5% , 97%)		

Figure B8. Additional statistics for the single-article *p*-curve for the number 6 of the top 10 (i.e., Mischel & Ebbesen, 1970).

Number 7 of the Top 10

The number 7 of the top 10 is the second study reported in the paper *Is Empathy-Induced Helping Due to Self-Other Merging* (Batson et al., 1997). Table B7 shows the score on each of the eight selected RDF for the number 7 paper.

Table B7

Coding paper nr. 7 of the top 10 (i.e., Batson et al., 1997)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "Two experiments tested the idea that empathy-induced helping is due to self-other merging." (Batson et al., 1997, p. 495) "Insofar as we know, there is as yet no clear evidence, pro or con, for the claim that empathy-induced helping is due to self-other merging. Therefore, we sought to test this claim, which we called the empathy-merging hypothesis, in each of two experiments by first replicating the empathy-helping relationship and then determining whether this relationship was due to self-other merging." (Batson et al., 1997, p. 497) "For each experiment, we predicted first that participants in the high-empathy condition, who were asked to imagine Katie 's feelings, would report more empathy for her than would participants in the low-empathy condition, who were asked to remain objective. We predicted second that participants in the high-empathy

		condition would offer more help to Katie than would participants in the low-empathy condition. If, however, shared group membership is a necessary condition for empathy, then these predictions should have been supported only under shared group membership. Support for these first two predictions would replicate the empathy-helping relationship and create the necessary conditions for testing competing hypotheses concerning the effect of inducing empathy on self-other perceptions and the effect of these perceptions on helping. The empathy-merging hypothesis predicted increased self-other merging in the high-empathy condition relative to the low and predicted that this merging would mediate the empathy-helping relationship. In contrast, the empathy-altruism hypothesis predicted increased empathy and helping in the high-empathy condition that was not mediated by self-other merging." (Batson et al., 1997, p. 498) "To test the idea that empathy-induced helping is due to self-other merging, we conducted two experiments. In each, we (a) presented participants with a need situation used in several studies demonstrating the empathy-helping relationship and (b) introduced a standard perspective-taking manipulation of empathy. Also, we manipulated shared group membership to check the generality of the empathy-helping relationship." (Batson et al., 1997, p. 506)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"It was not necessary to exclude anyone because of suspicion." (Batson et al., 1997, p. 503)
Sample size (predetermined or not).	3	"Participants for Experiment 2 were 60 general psychology students at the University of Kansas (40 men, 20 women), receiving partial credit toward a course requirement. Using a randomized block procedure, we assigned 10 men and 5 women to each cell of the 2 (empathy) x 2 (group membership) design." (Batson et al., 1997, p. 503)

Sharing/Openness (i.e., materials, data, code).	2	Materials: "Then the announcer interviewed Katie. She described her situation in these words:" (Batson et al., 1997, p. 499)
Using covariates and reporting the results with and without the covariates.	3	The seventh footnote: "The only reliable gender effect in Experiment 2 was a main effect on self-reported empathy, to be described next. Therefore, gender was not included as a factor in other reported analyses. We also found an effect of experimenter on helping. Although the pattern of helping across conditions was the same for all three experimenters, participants run by one male experimenter helped less than participants run by the other male and the female experimenter. Because this experimenter effect did not interact with the manipulations, we made no adjustment for it in the reported analyses." (Batson et al., 1997, p. 503)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"The design for each experiment was a 2 (empathy: low vs. high) x 2 (group membership: unshared vs. shared) factorial. After encountering a young woman in need, participants were given an unexpected chance to volunteer to help her. They also completed the three measures of self-other merging." (Batson et al., 1997, p. 498)
Fallacious interpretation of (lack of) statistical significance.	2	The fourth footnote concerning the first study, but also applies to the second study?: "For simplicity, all statistical tests are reported two-tailed, even for directional predictions." (Batson et al., 1997, p. 500) "In spite of the significantly higher IOS scores in the high-empathy condition, a path analysis using EQS revealed that IOS scores could not account for the empathy-helping relationship." (Batson et al., 1997, p. 505) "The empathy-merging hypothesis made a clear prediction that the conditions inducing empathy would produce increased self-other merging and that this merging would account for the empathy-helping relationship. The empathy-altruism hypothesis disagreed, predicting increased empathy and helping in the high empathy condition that could not be accounted for by self-other merging. Given that the empathy-altruism prediction was for no effect of merging, it is not appropriate to consider the present research a test of that hypothesis. All we can say is that our results are entirely consistent with the empathy-altruism hypothesis. Support for that hypothesis must come

		from research designed to provide direct tests. In the past 20 years, over 25 experiments have been so designed, and the support they provide is extensive (see Batson, 1991, for a review).” (Batson et al., 1997, p. 508) “Our results failed to support the idea that empathy-induced helping is due to self-other merging, leading us to answer a tentative no to the question with which we began. Merging of self and other into a psychological "one" does not seem to be the reason that empathy increases helping. In retrospect, we should perhaps have known that phrases such as "two shall become one," "self-expansion," "including other in the self," and "self-other merging" are best taken metaphorically rather than literally, at least when applied to empathy.” (Batson et al., 1997, p. 508)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	“Perception of Katie's need. As in Experiment 1, participants in all four cells of Experiment 2 perceived Katie's need to be great (cell means ranged from 8.00 to 8.53, overall M = 8.20), with no reliable difference among the cells, $F(3, 56) = 0.57$, $p > .50$. Effectiveness of the empathy manipulation. Also as in Experiment 1, results indicated that the empathy manipulation was effective. Participants in the low-empathy condition reported more concentration on being objective while listening to the broadcast (M = 7.73) than did participants in the high-empathy condition (M = 5.63), $F(1, 56)$ = 15.50, $p < .0005$. Conversely, participants in the high-empathy condition reported more concentration on Katie's feelings (M = 8.00) than did participants in the low-empathy condition (M = 3.60), $F(1, 56) = 101.81$, $p < .0005$. For neither measure did the main effect for group membership or the interaction approach significance (all Fs < 1.0).[7] Once again, to assess the effectiveness of the empathy manipulation in inducing empathic feelings for Katie, we created an index by averaging responses to the six empathy adjectives: sympathetic, compassionate, softhearted, warm, tender, and moved (Cronbach's a = .86). Scores on this 7-point empathy index (1 = not at all, 7 = extremely) were higher in the high-empathy condition (M = 4.98) than in the low-empathy condition (M = 3.48), $F(1, 52) = 24.55$, $p < .0005$ (see Row 1 of Table 2). Neither the group membership main effect nor the Empathy x Group Membership interaction approached significance (F

< 1.0). As has sometimes been found in the past, there was a main effect of gender of participant, with women reporting more empathy (M = 4.80) than men (M = 3.95), $F(1, 52) = 8.82$, $p < .005$. There were no reliable interactions between gender of participant and either experimental manipulation, and the difference between low- and high-empathy conditions was reliable for both men and women (ps < .001 and .02, respectively). We concluded that the empathy manipulation was effective in inducing empathic feelings for Katie among both men and women." (Batson et al., 1997, p. 503-504)

"An ANOVA on the continuous measure revealed only one reliable effect, a highly significant main effect for empathy, $F(1, 56) = 12.36$, $p < .01$. The group membership main effect was not reliable, $F(1, 56) = 2.27$, $p > .10$; and the interaction did not approach significance ($F < 1.0$). Participants in the high-empathy condition helped more (M = .67) than did participants in the low-empathy condition (M = .20), and participants in the shared group membership condition helped somewhat more (.53) than participants in the unshared group membership condition (.33, see Row 2 of Table 2). An ANOVA following arcsin transformation on the dichotomous measure produced exactly the same pattern of results. The empathy main effect was highly significant, $\chi^2(1, N = 60) = 11.03$, $p < .001$; the group membership main effect was not significant, $\chi^2(1, N = 60) = 1.40$, $p > .20$; and the interaction did not approach significance, x1 < 1.0 (see Row 3 of Table 2). The correlation between self-reported empathy and helping was highly significant for each helping measure (both rs(58) = .50, ps < .0005). Overall, these results indicated that we once again successfully replicated the empathy-helping relationship; indeed, replication was even clearer in Experiment 2 than in Experiment 1, possibly because the effect of the empathy manipulation on self-reported empathy was stronger." (Batson et al., 1997, p. 504)

"Inclusion of other in self.

Participants in the high-empathy condition again scored higher on the IOS scale (M = 2.93) than did participants in the low-empathy condition (M = 2.07); in Experiment 2 this difference, which was only marginal in Experiment 1, was reliable, $F(1, 56) = 5.23$, $p < .02$ (see Row 5 of Table 2). The group membership main effect did not approach significance ($F < 1.0$), but the Empathy x Group

Membership interaction was significant, $F(1, 56) = 4.88$, $p < .04$. This interaction was produced by a much larger difference between the means of the low- and high-empathy conditions in the unshared group membership condition (difference = 1.67) than in the shared group membership condition (difference = 0.06)." (Batson et al., 1997, p. 505)

"To check this possibility more directly, in Experiment 2 we included a measure of care about Katie's welfare. As expected, scores on this measure were significantly positively correlated with scores on the IOS scale, $r(58) = .30$, $p < .025$. In an analysis of covariance, controlling for scores on the care measure, the empathy main effect on IOS scores was no longer significant ($p > .10$); only the Empathy x Group Membership interaction remained, $F(1, 55) = 4.53$, $p < .04$. These results suggested that in the present context, IOS scores did indeed reflect, at least in part, care for Katie's welfare. Difference in perceived attributes of self and Katie. Rows 6 and 7 of Table 2 present mean absolute differences between ratings of self and Katie in Experiment 2, first for the nonrelevant attributes, then for the relevant. Once again, a 2 x 2 x 2 mixed-factor ANOVA revealed that overall, participants perceived greater difference between themselves and Katie on the relevant attributes ($M = 2.89$) than on the nonrelevant attributes ($M = 1.65$), $F(1, 56) = 79.77$, $p < .0005$, with no reliable effects of the experimental manipulations (either main effects or interactions). This difference across the measures again indicated more, not less, differentiation between self and Katie on relevant as opposed to nonrelevant attributes, even among high-empathy participants. Separate 2 x 2 between-group ANOVAs revealed that on the nonrelevant attributes, participants in the high-empathy condition perceived the same difference between themselves and Katie ($M = 1.65$) as did participants in the low-empathy condition ($M = 1.65$), $F(1, 56) = 0.00$, *n*s; other *F*s < 1.20. On the relevant attributes, participants in the high-empathy condition perceived slightly greater difference between themselves and Katie ($M = 2.92$) than did participants in the low-empathy condition ($M = 2.86$); in Experiment 2 (unlike Experiment 1) this effect did not approach statistical reliability (all *F*s < 1.85)." (Batson et al., 1997, p. 505)

In Figure B9, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(3, 56) = 0.57$; $F(1, 56) = 15.50$; $F(1, 56) = 101.81$; $F(1, 52) = 24.55$; $F(1, 52) = 8.82$; $F(1, 56) = 12.36$; $F(1, 56) = 2.27$; $\chi^2(1) = 11.03$; $\chi^2(1) = 1.40$; $F(1, 56) = 5.23$; $F(1, 56) = 79.77$; $F(1, 56) = 0.00$.

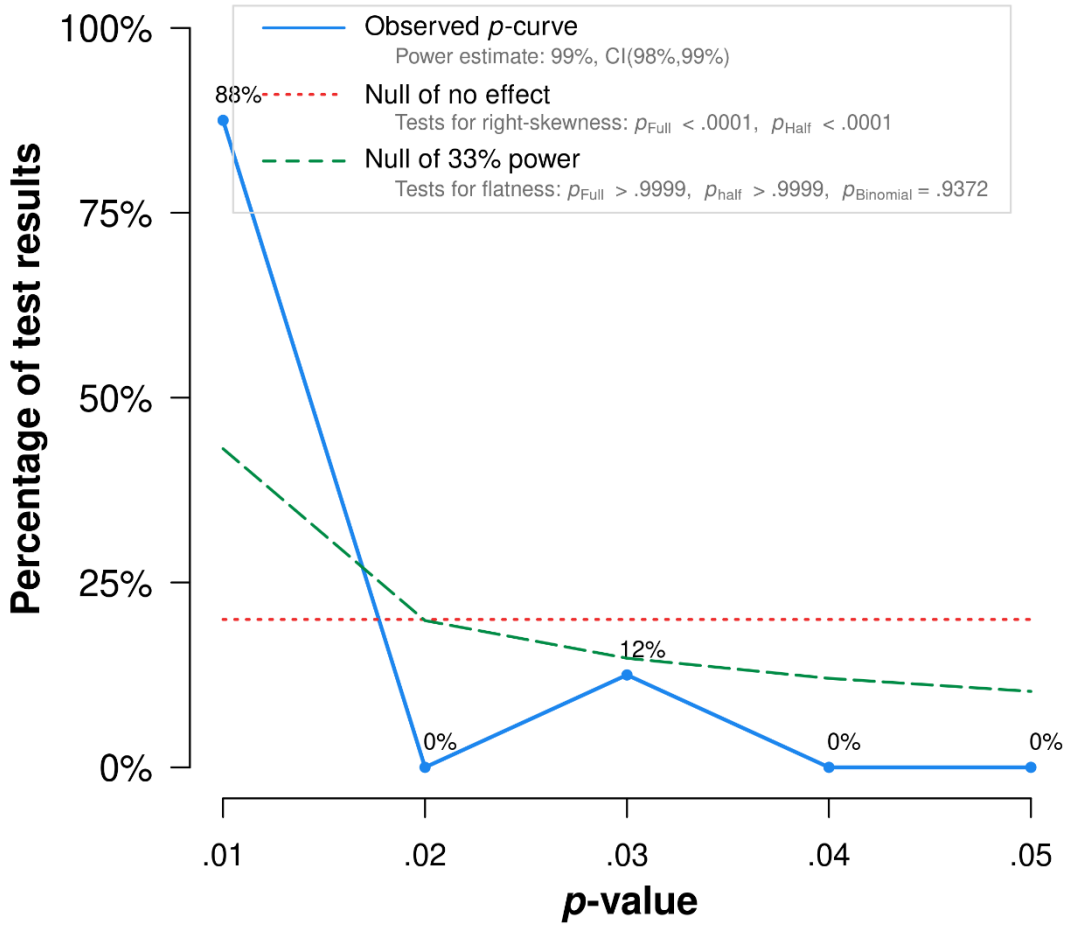


Figure B9. The single-article *p*-curve for the number 7 of the top 10 (i.e., Batson et al., 1997).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B10, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed (*p*s < .1), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the

33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B10, the 33% power test is $p > .9999$ for the full *p*-curve, for the half p-curve is $p > .9999$, and for the binomial 33% power test is $p = .9372$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.0352	*Z*=-8.93, *p*<.0001	*Z*=-8.94, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*=.9372	*Z*=6.51, *p*>.9999	*Z*=7.99, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (98% , 99%)		

Figure B10. The Additional statistics for the single-article *p*-curve for the number 7 of the top 10 (i.e., Batson et al., 1997).

Number 8 of the Top 10

The number 8 of the top 10 is the first study reported in the paper *When Approach Motivation and Behavioral Inhibition Collide: Behavior Regulation Through Stimulus Devaluation* (Veling et al., 2008). Table B8 shows the score on each of the eight selected RDF for the number 8 paper.

Table B8

Coding paper nr. 8 of the top 10 (i.e., Veling et al., 2008)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "In Experiment 1, using highly positive pictures as stimuli, we hypothesized overall lower attractiveness ratings to no-go stimuli compared to both go and new stimuli." (Veling et al., 2008, p. 1014) "In the present research we do not intend to study all implications of BSI theory, but we aim to test one specific hypothesis. Specifically, we aim to show that presentation of a positive stimulus together with a cue that signals that a response should be withheld, leads to devaluation of the positive stimulus. Furthermore, we expect that such inhibition induced devaluation occurs only with

positive stimuli and not with neutral and negative stimuli, albeit for different reasons: In the case of neutral stimuli because there is no response tendency in the first place (and hence withholding a response requires no inhibition), and in the case of negative stimuli because behavioral inhibition is not necessarily inconsistent with negative stimuli, and negative stimuli are more resistant to affective tuning." (Veling et al., 2008, p. 1014)

"We argue that in conflicting situations in which a stimulus is positive (e.g. you see a big glass of beer) while approach is undesirable (e.g. it is not yours) inhibition of the approach reaction will lead to devaluation of the positive stimulus. We tested this prediction in three experiments. Specifically, we tested whether behavioral inhibition elicited by a contextual cue in the presence of a positive stimulus results in devaluation of this stimulus." (Veling et al., 2008, p. 1013)

"Accordingly, we propose that whenever a response conflict arises between stimuli that trigger an approach reaction and cues that signal that approach is unwanted, behavioral inhibition and the stimuli interact, resulting in adaptive tuning of the valence of stimuli. We call this the Behavior Stimulus Interaction (BSI) theory. This tuning is the result of two interacting processes. More specifically, whenever a positive stimulus is encountered the approach system ensures that we get ready to respond. Because affective information is processed faster than other aspects of stimuli (see above) this approach tendency is always activated first. Next, the demands of the situation are processed. In circumstances where situational cues signal that approach towards the stimulus is unwanted, a response conflict is detected and the response will be inhibited. To solve this conflict then, the positive stimulus is devalued (i.e. negative affect is attached to it) to release the approach tendency, and tune its valence in line with the demands of the situation. As a result, the unwanted stimulus will be evaluated as less positive when it is subsequently encountered compared to a stimulus that did not give rise to a response conflict. (Of course, it may be that under some circumstances, e.g. when the stimulus becomes available again, the devaluation is cancelled.) The process just outlined may be functional because devaluation resulting from inhibition of the approach tendency ensures that a specific positive stimulus that first prompted a

		behavioral approach tendency will stop doing so, leaving room for other stimuli to take over guidance of behavior (Aarts et al., 2007). It is important to note that BSI theory pertains to inhibition of approach behavior, and not to avoidance or withdrawal behavior." (Veling et al., 2008, p. 1014)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	"Finally, we asked participants to type in what they thought to be the idea behind the experiment. In all experiments, none of the participants guessed the hypothesis of the study." (Veling et al., 2008, p. 1015)
Sample size (predetermined or not).	3	"Experiment 1 included 33 participants. In all experiments participants were students from Radboud University Nijmegen and received 1 euro (approximately $1.40) for their participation." (Veling et al., 2008, p. 1014)
Sharing/Openness (i.e., materials, data, code).	3	Footnote 1: "The IAPS picture identification numbers of the positive stimuli used in Experiments 1, 2, and 3 are 1440, 1460, 1750, 5000, 5010, 5200, 5780, 5700, 5982, 5830, 5760, and 8190." (Veling et al., 2008, p. 1015) However, the images are closed off "for access only to academic researchers to be used only in basic and health research projects" (https://imotions.com/blog/iaps-international-affective-picture-system/)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"We employed a 3 (stimulus status: go, no-go, new) one factorial within subjects design." (Veling et al., 2008, p. 1014-1015)

Fallacious interpretation of (lack of) statistical significance.	0	"results of three experiments show that refraining from responding to stimuli results in devaluation of these stimuli, but only when these stimuli are positive. These findings suggest automatic behavior-regulation, in terms of devaluation of positive stimuli, in situations in which environmental cues triggering approach (because of the positive valence of the stimulus) run counter to situational demands (cues that elicit behavioral inhibition)." (Veling et al., 2008, p. 1013) "These results are in line with BSI theory: Specific positive stimuli are devalued when situational cues have repeatedly elicited behavioral inhibition upon encountering these stimuli. The fact that the no-go stimuli were rated as less attractive compared to the new stimuli is especially indicative of devaluation. The result that merely not responding to specific stimuli in a go-no-go task causes devaluation of these specific stimuli compared to new stimuli in a subsequent evaluation task is a novel finding. Nonetheless, an even more direct test of the theory would be to show that valence of stimuli and behavior interact, so that only behavioral inhibition to positive stimuli and not to neutral stimuli would result in devaluation of the no-go stimuli. This is what we aimed to show in Experiment 2 by including pictures in the go/no-go task that are of neutral valence." (Veling et al., 2008, p. 1015) "In three experiments we showed that consistently not responding to positive stimuli leads to devaluation of these stimuli compared to stimuli to which a response was required, and compared to new stimuli." (Veling et al., 2008, p. 1017)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	3	"To test whether repeated pairing of specific stimuli (i.e. pictures) with a no-go response would cause devaluation of these no-go stimuli compared to both new stimuli and go stimuli we performed repeated measures analysis of variance (ANOVA) with one factor (stimulus status: go, no-go, new). This analysis revealed the predicted effect of stimulus status, $F(2, 64) = 3.33$, $p < .05$, partial $\eta^2 = .09$. Simple effect analyses revealed that participants evaluated no-go stimuli ($M = 5.33$, $SD = 0.85$) reliably lower than both go stimuli ($M = 5.77$, $SD = 0.91$), and new stimuli ($M = 5.81$, $SD = 1.11$), respective comparisons $F(1, 32) = 4.59$, $p < .05$, $\eta^2 = .13$ and $F(1, 32) = 6.20$, $p < .05$, $\eta^2 = .16$. There was no reliable difference between go and new stimuli $F(1, 32) < 1.^2$" (Veling et al., 2008, p. 1015)

In Figure B11, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(2, 64) = 3.33$; $F(1, 32) = 4.59$; $F(1, 32) = 6.20$.

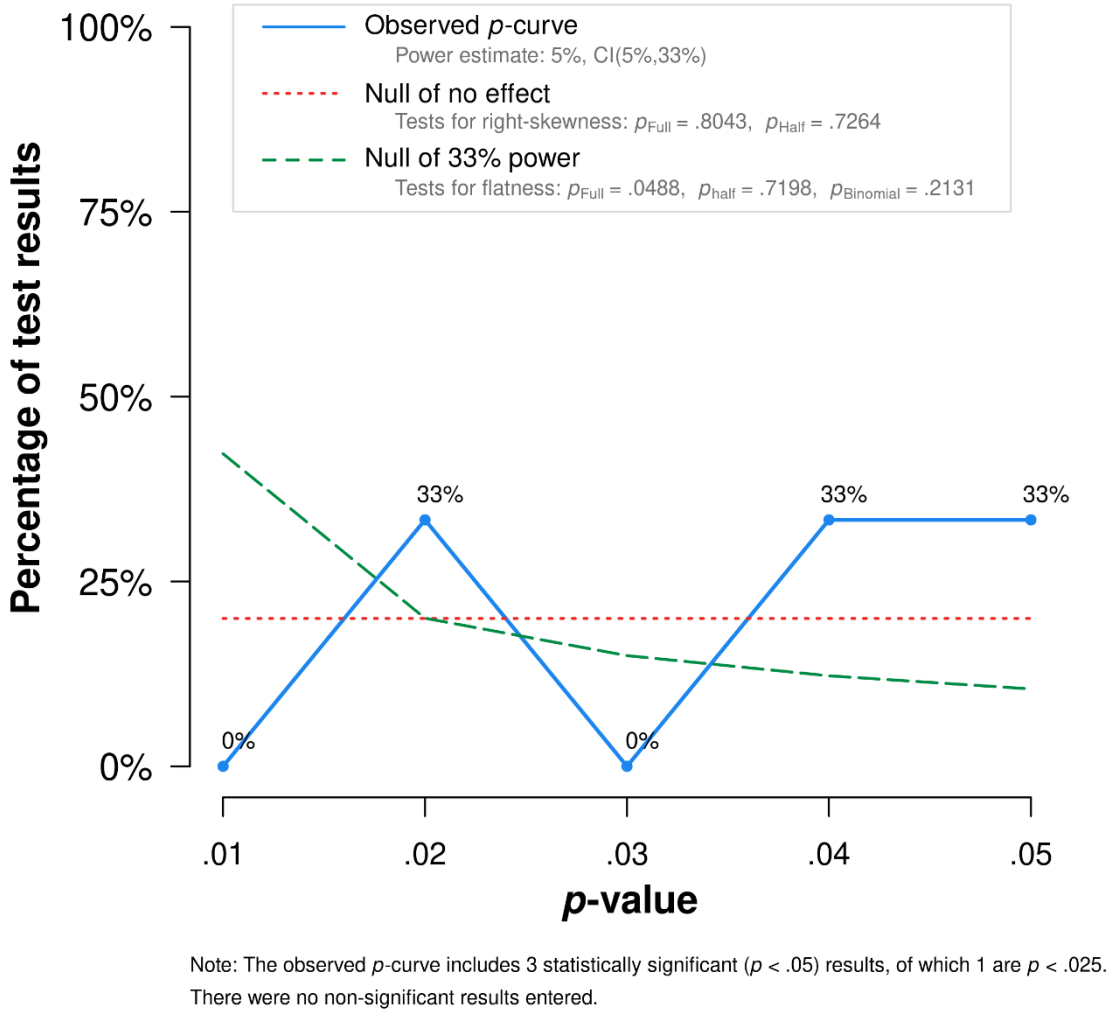


Figure B11. The single-article *p*-curve for the number 8 of the top 10 (i.e., Veling et al., 2008).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B12, not only is the half *p*-curve test ($p = .7264$) not significantly right-skewed ($p < .05$), but also both the half ($p = .7264$) and full test ($p = .8043$) are not significantly right-skewed ($ps < .1$), which implies that the study does not contain evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power

test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B12, the 33% power test for the half *p*-curve is $p = .7198$, and for the binomial 33% power test is $p = .2131$. However, $p = .0488$ for the full *p*-curve, which indicates that evidential value in the study is inadequate or absent." ("*P*-curve results app 4.06," 2017) Furthermore, there is even comment in bold text after running this *p*-curve in the app: "direct replications of the submitted studies are not expected to succeed." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	p=.875	Z=0.86, p=.8043	Z=0.6, p=.7264
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p=.2131	Z=-1.66, p=.0488	Z=0.58, p=.7198
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 5% 90% Confidence interval: (5% , 33%)		

Figure B12. Additional statistics for the single-article *p*-curve for the number 8 of the top 10 (i.e., Veling et al., 2008).

Number 9 of the Top 10

The number 9 of the top 10 is the first study reported in the paper *The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis* (Lodge & Taber, 2005). Table B9 shows the score on each of the eight selected RDF for the number 9 paper.

Table B9

Coding paper nr. 9 of the top 10 (i.e., Lodge & Taber, 2005)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	"In each of the studies, we hypothesize that reaction times will be faster for affectively congruent prime-target concepts (pos/pos and neg/neg) than for incongruent pairs (neg/pos and pos/neg). This is the basic hot cognition hypothesis. Critical to the hot cognition postulate is that one's feelings are triggered automatically on the mere presentation of the concept; accordingly, the predicted facilitation and inhibition effects should only show up in the short SOA condition when priming activation is at peak. Operationally, our most basic hypothesis is

represented by the three-way interaction, SOA x prime valence x target valence. Note that we have no expectations about differential effects for negative or positive primes or targets, but only about the affective congruence of prime-target pairs. These projected analyses will be broken down by sophistication (a between subjects correlate) and attitude strength (within subjects). In general, we predict that political sophisticates and those with strong attitudes would be most likely to have formed online affective links for all of the political objects we use as primes and so we expect stronger results for sophisticates than for unsophisticates and for primes that evoke strong attitudes. Finally, the basic reason given for the expectation that groups and issues are less likely to be linked to evaluative affect is that attitudes toward these objects are thought to consciously evoke pro and con considerations and consequently be more ambivalent than are attitudes toward persons. Therefore, in addition to comparing hot cognition for the three prime types, we will directly test the underlying contention that implicit attitudes should be weaker for ambivalent primes." (Lodge & Taber, 2005, p. 467)

"three experimental tests of the "hot cognition" hypothesis, which posits that all sociopolitical concepts that have been evaluated in the past are affectively charged and that this affective charge is automatically activated within milliseconds on mere exposure to the concept, appreciably faster than conscious appraisal of the object" (Lodge & Taber, 2005, p. 455)

"In this paper we report the results of three experimental studies testing a central postulate of our dual-process model of motivated political reasoning (Lodge & Taber, 2000; Taber & Lodge, 2001; Taber, Lodge, & Glather, 2001). This theory of motivated reasoning starts with the hot cognition hypothesis (Abelson, 1963), the claim that all sociopolitical concepts are affect laden (Bargh, 1994, 1997; Fazio & Williams, 1986; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Fiske, 1982; Lodge & Stroh, 1993; Lodge, McGraw, & Stroh, 1993; McGraw, Lodge, & Stroh, 1990; Morris, Squires, Taber, & Lodge, 2003). All political leaders, groups, issues, symbols, and ideas thought about and evaluated in the past become affectively charged—positively or negatively—and this affect is linked directly to the concept in long-

		term memory. This evaluative tally, moreover, comes automatically and inescapably to mind upon presentation of the associated object, thereby signaling its affective coloration (what Clore & Isbell [2001] call the "how-do-I-feel heuristic?" and what Sniderman, Brody, & Tetlock [1991] call the "likability heuristic"). At the moment one realizes that the letters B-U-S-H in a news headline refer to the President and not to a plant, one's affect toward "W" Bush comes to mind along with his strongest cognitive associations." (Lodge & Taber, 2005, p. 456) "The studies reported here directly test the hot cognition question: are attitudes toward political leaders, groups, and issues evoked automatically or do they require a more effortful—and time-consuming—process of evaluative integration?" (Lodge & Taber, 2005, p. 457)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"By their nature, reaction time data are highly positively skewed, and this skewness can affect group means in the analysis of variance. To correct for positive skewness in our data (Study 1, skewness = 3.59; Study 2 = 3.74; Study 3 = 2.83), we subjected the raw reaction time data to a natural log transformation (Bargh & Chartrand, 2000; Fazio, 1990, 1993). All statistical results reported below are computed on these natural log transformed reaction time data; it is worth noting, however, that the overall pattern of results emerges with or without this transformation. In addition, we eliminated trials involving targets that had been incorrectly rated in the survey (e.g., someone might say that "miserable" was a good thing, in which case we excluded the trials for that subject in which miserable was the target; .04% of trials across the three studies), and we eliminated trials in which there was an incorrect response to the target on the RT (error rate of 5% across the three studies)." (Lodge & Taber, 2005, p. 466-467)
Sample size (predetermined or not).	3	"Undergraduate students in introductory political science courses at Stony Brook University received extra credit for their participation: Study 1, N = 80" (Lodge & Taber, 2005, p. 463)
Sharing/Openness (i.e., materials, data, code).	2	Materials in Table 1: "Primes and Targets" (Lodge & Taber, 2005, p. 464)

Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"Studies 1 and 2 were two (SOA, long vs. short) x 2 (prime valence, positive vs. negative) x 2 (target valence, positive vs. negative) mixed model designs with repeated measures on prime and target valence" (Lodge & Taber, 2005, p. 467)
Fallacious interpretation of (lack of) statistical significance.	3	Calling an effect strong based on statistical significance: "Even semantically unrelated affective concepts (e.g., "sunshine," "cancer") have a strong effect on the evaluation of political leaders, groups, and issues." (Lodge & Taber, 2005, p. 455)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"Looking first at the basic prediction for Study 1 for all political primes, we find strong support for the hypothesized three way interaction of SOA, prime, and target, $F(1, 78) = 14.29$, $p < .001$, with no significant main effects. This result is captured in Figure 4a, which contrasts the basic expected pattern of facilitation and inhibition effects at short SOA, with no facilitation/inhibition effects at long SOA. Follow up contrasts confirm the apparent pattern in Figure 4a: under short SOA, responses to negative targets are significantly faster when preceded by negative primes, $t(45) = 2.02$, $p = .025$ (one-tailed), while positive primes elicit faster response times when paired with positive targets, $t(44) = 2.26$, $p = .02$. As predicted, similar contrasts for long SOA failed to reach significance. (To reduce redundancy, we will limit the remaining figures to the short SOA condition, though we will continue to report the full interactions in text.)" (Lodge & Taber, 2005, p. 468)

In Figure B13, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(1, 78) = 14.29$; $t(45) = 2.02$; $t(44) = 2.26$.

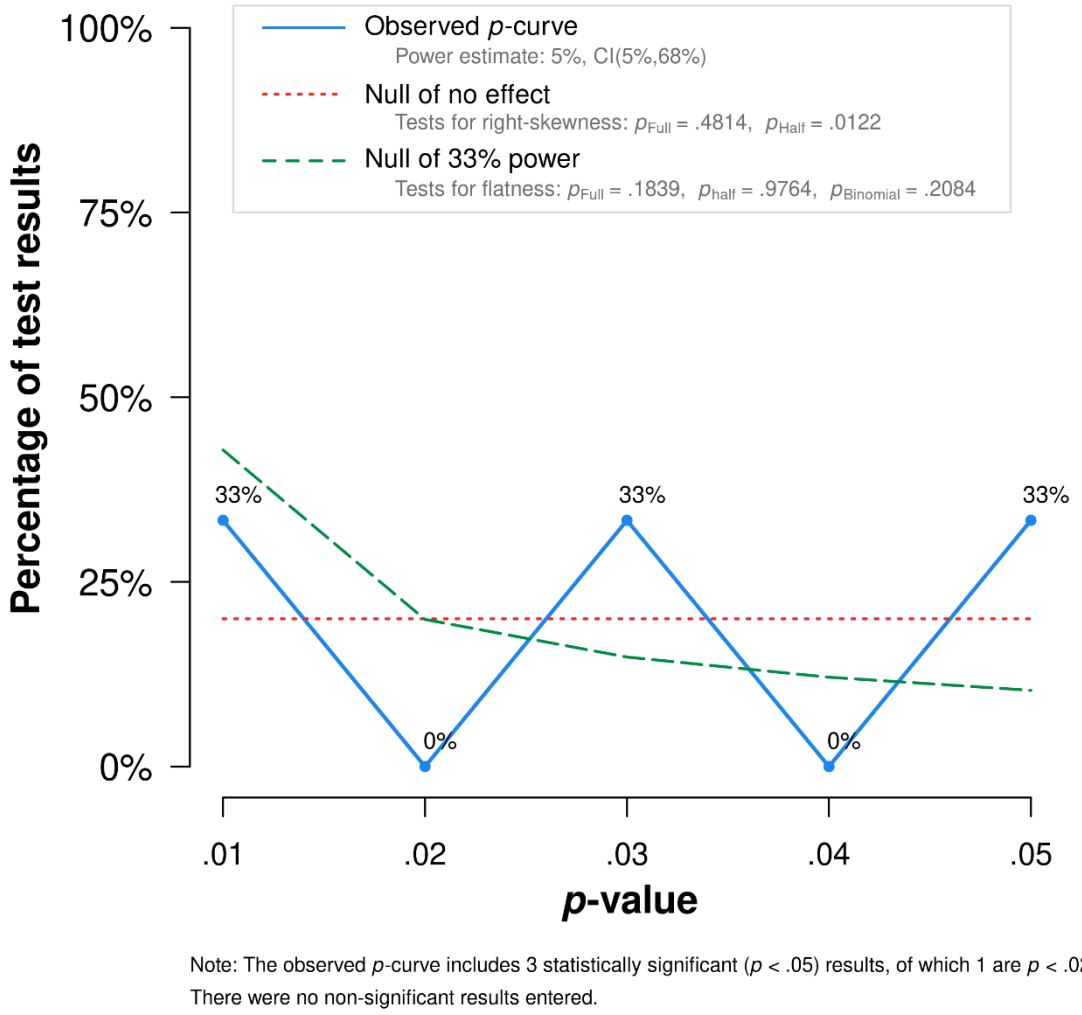


Figure B13. The single-article *p*-curve for the number 9 of the top 10 (i.e., Lodge & Taber, 2005).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B14, the half *p*-curve test ($p = .0122$) is significantly right-skewed ($p < .05$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B14, the 33% power test is $p = .1839$ for the full *p*-curve, for the half p-curve is $p = .9764$, and for the binomial 33%

power test is $p = .2084$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	Continuous Test *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. *(Right skew)*	p=.875	Z=-0.05, p=.4814	Z=-2.25, p=.0122
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p=.2084	Z=-0.9, p=.1839	Z=1.98, p=.9764
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 5% 90% Confidence interval: (5% , 68%)		

Figure B14. Additional statistics for the single-article *p*-curve for the number 9 of the top 10 (i.e., Lodge & Taber, 2005).

Number 10 of the Top 10

The number 10 of the top 10 is the first study reported in the paper *Affective and Physiological Responses to the Suffering of Others: Compassion and Vagal Activity* (Stellar et al., 2015). Table B10 shows the score on each of the eight selected RDF for the number 10 paper.

Table B10

Coding paper nr. 10 of the top 10 (i.e., Stellar et al., 2015)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "In keeping with the caretaking origins of compassion and the types of behaviors it motivates, in the present investigation, we tested the hypothesis that compassion will be accompanied by greater parasympathetic activity than will neutral states or closely related positive and prosocial states. We measured parasympathetic activity via the vagus nerve, which we operationalized as respiratory sinus arrhythmia (RSA). This hypothesis has its precedent in theoretical arguments about the evolution of the vagus nerve (Porges, 2001) and was anticipated by select findings, to which we now turn." (Stellar et al., 2015, p. 573) "In the present studies, we explore the peripheral physiological changes associated with the experience of compassion. Guided by long-standing theoretical claims, we propose that compassion is associated with activation in the parasympathetic

		autonomic nervous system through the vagus nerve. Across 4 studies, participants witnessed others suffer while we recorded physiological measures, including heart rate, respiration, skin conductance, and a measure of vagal activity called respiratory sinus arrhythmia (RSA). Participants exhibited greater RSA during the compassion induction compared with a neutral control (Study 1)," (Stellar et al., 2015, p. 572) "Study 1 primarily aimed to demonstrate that (a) inducing compassion by showing others who are suffering elicits greater RSA than does a nonemotional control, and (b) differences in RSA between the compassion and neutral state inductions predict self-reports of compassion. We also explored whether RSA during the nonemotional control predicted any of our measures. We measured and controlled for individual differences in social desirability, because self-reports of compassion are likely to be influenced by self-presentation concerns that could obscure the relationship between the experience of compassion and RSA." (Stellar et al., 2015, p. 575)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	3	"Artifacts in the signal (e.g., due to coughing, sneezing, movement) were corrected manually; this was done to less than 5% of all data files." (Stellar et al., 2015, p. 575)
Sample size (predetermined or not).	3	"Fifty-one (20 male, 31 female) undergraduates from a large west coast U.S. university participated in this study for credit in a psychology course." (Stellar et al., 2015, p. 575)
Sharing/Openness (i.e., materials, data, code).	3	
Using covariates and reporting the results with and without the covariates.	0	No covariates.

Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"We used a linear mixed-model approach that treated the type of video (neutral or compassion inducing) as a within-subject factor. This statistical analysis allowed us to calculate differences in physiological activity within the same person over multiple emotion inductions while controlling for dependencies in the same person's data across time. Linear mixed models also allowed for the inclusion of respiration as a changing covariate, which was important, because respiratory rate was different for each emotion induction." (Stellar et al., 2015, p. 575-576) "In Study 1, we compared RSA during a compassion induction, in which targets discussed the death of their grandfather, to a nonemotional baseline. (…) Across the first three studies, we assessed the relationship between retrospective self-reports of compassion and differences in RSA between the compassion and comparison inductions." (Stellar et al., 2015, p. 575)
Fallacious interpretation of (lack of) statistical significance.	0	"The present investigation demonstrates that the experience of compassion, when encountering the suffering of others, leads to markedly greater vagal activity compared with neutral or other emotional states." (Stellar et al., 2015, p. 575)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"Participants exhibited high levels of compassion in response to the student coping with the death of her grandfather (M = 7.20, SD = 1.61). Participants reported significantly more compassion than the next two most highly elicited emotions, which were sadness, (M = 6.29, SD = 2.34), $t(50) = 3.93$, $p <$ .001, and warmth/tenderness, (M = 5.37, SD = 2.38), $t(50) = 7.29$, $p < .001$, which were the only two emotions with average self-reports greater than the midpoint of the scale." (Stellar et al., 2015, p. 575) "RSA was significantly higher during the compassion condition (M = 79.06 ms, SD = 25.59 ms) than the neutral condition (M = 69.40 ms, SD = 34.30 ms), $F(1, 50) = 7.12$, $p = .01$, d = 0.39. Emotions are rapid in onset and brief in duration (Ekman, 1992). Therefore, we also examined the first min and a half of RSA (the shortest acceptable duration to analyze RSA; Berntson et al., 1993) because we anticipated that participant's reactions to the compassion induction would be the strongest on initially encountering the target's suffering. We found that our effects were amplified when we examined this initial 1.5 min. RSA was

significantly higher in the compassion condition (*M* = 82.34 ms, *SD* = 34.02 ms) than the neutral condition (*M* = 70.16 ms, *SD* = 26.24 ms), F(1, 50) = 12.43, *p* = .001, *d* = 0.51.[2] Compassion-inducing stimuli elicited greater RSA than a nonemotional condition, and these effects were more pronounced in the initial stages of the stimulus presentation." (Stellar et al., 2015, p. 576)

The second footnote: "We also examined differences in RSA between conditions using repeated measures and found that our results in this and all subsequent studies exhibited the same pattern of significance as when analyzed using linear mixed models (*F*s ≥ 3.85, *p*s ≥ 05)." (Stellar et al., 2015, p. 576)

"Respiration rate was lower over the entire 4 min of the compassion video (*M* = 18.33, *SD* = 4.70) than it was in the neutral condition (*M* =19.33, *SD* = 4.43), *F*(1, 50) = 7.79, *p* = .007; *d* = 0.40. We conducted a mixed model with type of video as a within-subject factor and respiration rate as a changing covariate and still found that RSA was significantly higher in the compassion condition than in the neutral condition over the entire 4 min, *F*(1, 53) = 4.61, *p* = .04. Within the 1st minute and a half of the video, there were no significant differences in respiration rate between the neutral (*M* = 19.23, *SD* = 3.78) and compassion conditions (*M* = 18.18, *SD* = 4.87), *F*(1, 50) = 2.67, *p* = .11. Nevertheless, when we treated respiration rate as a covariate, we still observed that RSA was significantly higher in the compassion condition than in the neutral condition, *F*(1, 51) = 9.61, *p* = .003. Surprisingly, we did not find significant differences in heart rate, which often accompany compassionate responses, during the compassion condition (*M* = 79.11, *SD* = 8.54) compared with the neutral condition (*M* = 78.37, *SD* = 8.31), *F*(1, 50) = 2.80, *p* = .10. Overall, we found a pattern of greater RSA during the compassion induction compared with a nonemotional control. As we predicted, these effects were more pronounced in the 1st minute and a half of the video, during which individuals first encountered the suffering." (Stellar et al., 2015, p. 576)

In Figure B15, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve results app 4.06," 2017): $t(50) = 3.93$; $t(50) = 7.29$; $F(1, 50) = 7.12$; $F(1, 50) = 12.43$; $F(1, 50) = 7.79$; $F(1, 53) = 4.61$; $F(1, 50) = 2.67$; $F(1, 51) = 9.61$; $F(1, 50) = 2.80$.

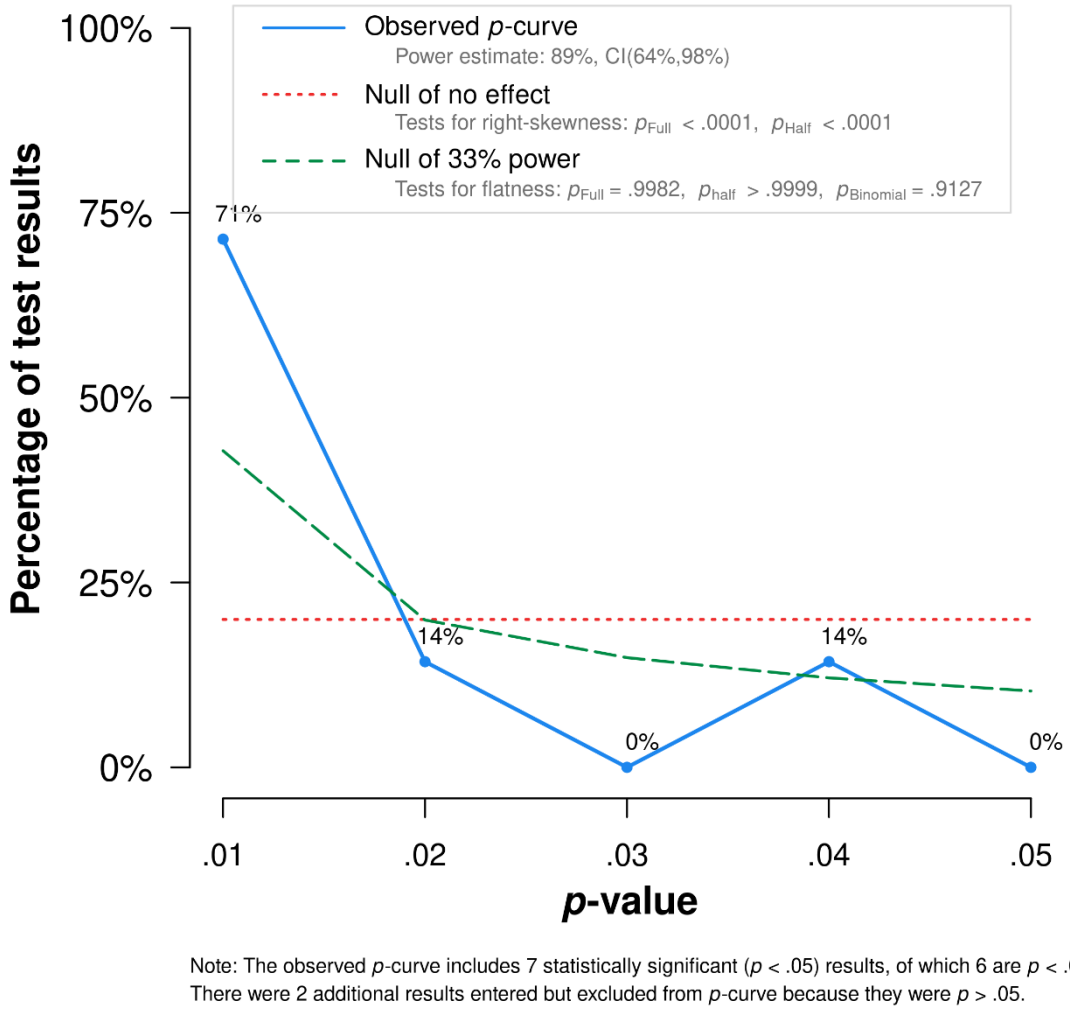


Figure B15. The single-article *p*-curve for the number 9 of the top 10 (i.e., Stellar et al., 2015).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure B16, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed (*p*s < .1), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power

test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure B16, the 33% power test is $p = .9982$ for the full *p*-curve, for the half *p*-curve is $p > .9999$, and for the binomial 33% power test is $p = .9127$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.0625	*Z*=-4.84, *p*<.0001	*Z*=-4.59, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*=.9127	*Z*=2.91, *p*=.9982	*Z*=4.65, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 89% 90% Confidence interval: (64% , 98%)		

Figure B16. Additional statistics for the single-article *p*-curve for the number 9 of the top 10 (i.e., Stellar et al., 2015).

Appendix C. Scoring the Center 10 Studies on the RDF Checklist

In order to map the DFS of the center 10 studies, each of the ten studies is scored on RDF. This Appendix contains the scores for each study on each of the eight items on the RDF checklist.

Number 1 of the Center 10

The number 1 of the center 10 is the first study reported in the paper *A research experience for American Indian undergraduates: Utilizing an actor–partner interdependence model to examine the student–mentor dyad* (Griese et al., 2017). Table C1 shows the score on each of the eight selected RDF for the number 1 paper.

Table C1

Coding paper nr. 1 of the center 10 (i.e., Griese et al., 2017)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "The goal of the present study, therefore, is to examine the student and mentor dyads within the SURE program, an often overlooked or inadequately assessed relationship impacting student outcomes. In the current study an APIM will be used to examine the actor and partner effects present in the student–mentor relationship. As seen in Figures 1 through 4 depicting APIMs, there are various effects tested within the model (implying relationship, not causation). It is expected that students' and mentors' beliefs at baseline, or time one (T1) would be associated with their beliefs or behaviors postprogram, or time two (T2), indicating significant actor effects. It is hypothesized that in accounting for actor paths in the overall model, that mentor's beliefs about the importance of skills at T1 would significantly impact students self-rated skills at T2 (partner path). These findings would support the notion that while student's beliefs regarding their own skills are important, the impact of the mentor's beliefs regarding the importance of a skill is integral to the student–mentor relationship and subsequent student outcomes. We would expect that the mentor's beliefs at T1 would significantly impact student skills at T2 such that high mentor beliefs would be associated with high student outcomes or behaviors at T2. This relationship is expected for the following models: research process knowledge, innovation, autonomy, and academic

		resilience. We will also test the partner path between student self-rated skills at T1 and mentor ratings at T2. A significant path may indicate that students' beliefs about their own behaviors at the beginning would be exuded throughout the program, in turn impacting the mentors' perceptions of their skills post program." (Griese et al., 2017, p. 42) "The present study will examine the mentor–student dyad within a 10-week summer research experience for American Indian undergraduates." (Griese et al., 2017, p. 40) "The present study is part of a broader evaluation of the Summer Undergraduate Research Experience (SURE). SURE is a 10-week URE where students obtain hands-on experience in biomedical or behavioral research projects. AI undergraduate students interested in health care careers or health disparity research are recruited and matched with a mentor to work in an area of interest. Mentors are drawn from a research institute under a health care organization (lead organization), local Veteran's Affairs research arm, and one public university. Students attend weekly seminars focused on professional development (e.g., literature searches) and health disparities research (e.g., history of research with AI/Alaska Native). Throughout these experiences, students learn strategies for working within the research field, including problem-solving skills and general research knowledge. There are also scaffolded opportunities through which students are guided by their mentors and in the end are able to engage in the activity on their own; thus, building their sense of autonomy within research." (Griese et al., 2017, p. 41) "The goal of the current study was to examine the student–mentor dyad at the beginning and end of a 10-week summer research experience for American Indian undergraduates utilizing a series of actor–partner interdependence models within SEM." (Griese et al., 2017, p. 39)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants	0	No exclusions.

in analyses. Reporting on how to deal with outliers in an ad hoc manner.		
Sample size (predetermined or not).	3	"**Participants** In the present study, the sample included 26 students (19 women), with a mean age of 24 years (*SD* = 6.4, range 18–45). Race or ethnicity breakdown for the students was as follows: AI/AN only (50%), AI/AN and White (50%). There were 27 mentors (15 men) who took part in the SURE program over the four summers, including primary and secondary mentors, with a mean age of 42 years (*SD* = 12.7, range 22–68). Race or ethnicity breakdown for mentors was as follows: 89% White and 11% Asian." (Griese et al., 2017, p. 43) "Participants included 26 undergraduate interns (50% American Indian; 50% American Indian and White; *M* age = 24) and 27 mentors (89% White; *M* age = 47)." (Griese et al., 2017, p. 39)
Sharing/Openness (i.e., materials, data, code).	3	"Primary analyses tested APIM within Structural Equation Modeling (SEM) using Mplus software (v. 7.11; Muthén & Muthén, 2012)." (Griese et al., 2017, p. 44) "Pre- and postsurvey measures were designed to investigate the perceptions of students and mentors regarding their skills, character, and program experiences. All SURE students and mentors were invited via email to participate in pre- and postsurveys using the Survey Monkey online survey system. Data were collected in the summers of 2011–2014 at the beginning and end of the SURE program. At the onset of the program, program staff explained study procedures. Participants were not compensated for their participation. All study procedures were approved by Sanford Research Institutional Review Board (IRB) and University of South Dakota IRB. Baseline data was collected during the first week of the program, and postsurveys were collected in the last week of the program." (Griese et al., 2017, p. 42-43) "**Measures** **Student and mentor prompts.** Students and mentors were surveyed regarding similar behaviors and outcomes including research process knowledge, academic character, academic resilience, and innovation; however, there were

		differences in the prompts provided. (…) **Research process knowledge.** Student- and mentor-reported Research Process Knowledge was measured at T1 and T2 via 14 questions adapted from Kardash (2000). Example items included (…) **Innovation.** Student- and mentor-reported Innovation was measured at T1 and T2 via 3 questions adapted from Singer and Weiler (2009). Items included (…) **Autonomy.** Student- and mentor-reported Autonomy was measured at T1 and T2 via 4 questions adapted from Singer and Weiler (2009). Example items included (…) **Academic resiliency.** Student- and mentor-reported Academic Resiliency was measured at T1 and T2 via three questions adapted from Singer and Weiler (2009). Items included (…)" (Griese et al., 2017, p. 43)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	1	"Although both multilevel modeling and SEM strategies can be used to analyze APIM, the following guidelines from Kenny et al. (2006) were utilized here. An advantage of SEM is that the model in its entirety is estimated, allowing all variables that would otherwise have to be estimated separately, to be estimated simultaneously (Kenny et al., 2006). SEM is also preferred over pooled-regression approaches wherein homogeneity of variance (where the variance for the student and mentor are assumed to be equal) is assumed; SEM approaches to APIM do not require this assumption. Finally, APIM effects can be viewed simply as a path-analytic model and, therefore, are easily translatable. For this study, path analysis within an SEM framework was used to analyze APIMs (Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). The following models estimate all potential paths between student and mentor, including all actor and partner paths. Because of this, the APIMs tested here were saturated (or just-identified, $df = 0$) and fit statistics were, therefore, irrelevant and not reported." (Griese et al., 2017, p. 44-45)

Fallacious interpretation of (lack of) statistical significance.	2	"Given the nonsignificance of the path of interest (Mentor T1 to Student T2), these findings suggest that what the mentors believe regarding the importance of innovation at T1 is not significantly associated with student's innovation at T2." (Griese et al., 2017, p. 45-46) "**Student and Mentor Change** Descriptive analyses (*t* tests) indicated a change in student-rated skills from the beginning to the end of the SURE program. These findings suggest the important impact of the program on students' skill levels, indicating students significantly increased in their self-reported research process knowledge, autonomy, academic resilience, and innovation over the 10-week program. For the mentors, *t* tests indicated that, in general, they had high beliefs regarding the importance of all skills at T1; however, their beliefs about their student's ability to portray these skills at T2 were lower. For two of the measures, academic resilience and research knowledge skills, there was a significant decline from T1 to T2 for mentors. This decline may be due, in part, to mentors overestimating their expectations of the students at the beginning of the summer. Although it was important to the mentors that students gain high levels of each of the skills as indicated by their T1 scores, the T2 scores suggest that throughout the course of the 10-week program, mentors may have gained a more realistic understanding of the skills that could be attained over the short time period." (Griese et al., 2017, p. 47) "Findings indicated that in accounting for all potential paths between students and mentors, the partner path between mentor beliefs at the beginning of the program and students' skills related to autonomy ($\beta = .59$, $p = .01$) and academic resilience ($\beta = .44$, $p = .03$) at the end of the program were significant. These findings suggest the important impact of mentor beliefs on student outcomes, a relationship that should be adequately assessed and continue to be important focus of undergraduate research experiences." (Griese et al., 2017, p. 39)
Assessing the evidential value of a single article by judging the single-article *p*-curve	0	"Preliminary analyses included an examination of bivariate correlations (see Table 1). Means, *SD*s, and ranges are also reported along with results from paired-sample t tests (see Table 2) that examined

(Simonsohn et al., 2014).	the change in student-reported skills and mentor-reported perceptions at the beginning and end of the SURE program. Students. Paired-sample t tests indicated that students' self-reported research skills significantly changed from T1 to T2 for all of the measures examined; innovation ($t(20) = -4.24$, $p < .001$), academic resilience ($t(20) = -3.28$, $p < .01$), autonomy ($t(19) = -5.12$, $p < .001$), and research process knowledge ($t(20) = -3.29$, $p < .01$). These findings indicate that student reported behaviors significantly increased on all measured behaviors from the beginning of the program to the end. **Mentors.** Mentors' beliefs regarding the importance of a skill and perception of their student's ability at the end of the summer did not significantly change from T1 to T2 for innovation, academic resilience, and autonomy, indicating that mentor expectations did not deviate significantly from their reported student outcomes. There was a significant change in mentors' perceptions about students' research process knowledge ($t(20) = 2.55$, $p < .05$). However, this change was negative, indicating their beliefs regarding the importance of research process knowledge at T1 were higher than their perceptions regarding their student's abilities at T2." (Griese et al., 2017, p. 43-44)

In Figure C1, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $t(20) = -4.24$; $t(20) = -3.28$; $t(19) = -5.12$; $t(20) = -3.29$; $t(20) = 2.55$.

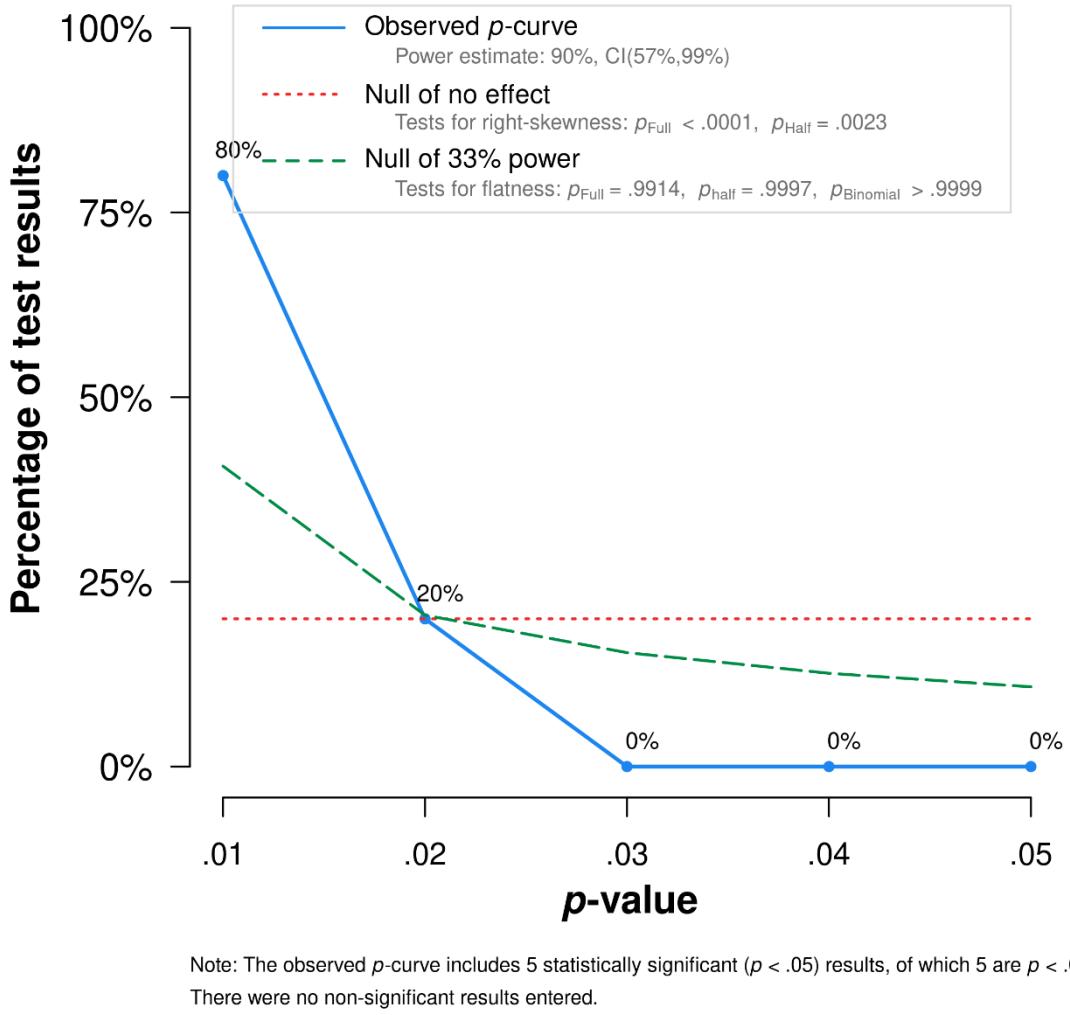


Figure C1. The single-article *p*-curve for the number 1 of the center 10 (i.e., Griese et al., 2017).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C2, not only is the half *p*-curve test ($p = .0023$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p = .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C2, the 33% power test is $p = .9914$ for the full *p*-curve, for the half *p*-curve is $p = .9997$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.0313	Z=-3.86, p=.0001	Z=-2.83, p=.0023
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p>.9999	Z=2.38, p=.9914	Z=3.43, p=.9997
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 90% 90% Confidence interval: (57% , 99%)		

Figure C2. Additional statistics for single-article p-curve for the number 1 of the center 10 (i.e., Griese et al., 2017).

Number 2 of the Center 10

The number 2 of the center 10 is the second study reported in the paper *Environmental Resources and the Posttreatment Functioning of Alcoholic Patients* (Bromet & Moos, 1977). Table C2 shows the score on each of the eight selected RDF for the number 2 paper.

Table C2

Coding paper nr. 2 of the center 10 (i.e., Bromet & Moos, 1977)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "In order to test this hypothesis" (Bromet & Moos, 1977, p. 335) "This paper examines the role of environmental resources in the posttreatment adjustment (drinking, psychological, and social functioning) of alcoholic patients who had previously been treated in residential alcoholism programs." (Bromet & Moos, 1977, p. 327-328) "The present research assessed the posttreatment functioning of alcoholic patients in relation to (I) the presence or absence of marital and occupational resources, and (2) if present, the type of social environment they provided." (Bromet & Moos, 1977, p. 326)
Exclusion of participants (how many, why, etc.).	1	"Patients in the study included approximately 80 percent of all admissions during a ten-month average period. Most nonparticipants dropped out

Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.		of their programs within a few days of admission." (Bromet & Moos, 1977, p. 328) "*Follow-up Data* *Posthospital functioning.* A follow-up evaluation, conducted approximately 6-8 months after discharge, was obtained from a questionnaire identical in content to the Background Information Form ad- ministered at intake. The 429 patients who completed the questionnaire comprised 87 percent of the original sample of 494 patients available for follow-up (excluding 11 patients who died during the follow-up period). There were no significant differences in background characteristics within any of the programs between patients included in the follow-up and those not participating." (Bromet & Moos, 1977, p. 329-330)
Sample size (predetermined or not).	3	"The sample was composed of 429 alcoholic patients drawn from five residential alcoholism programs. These facilities represent a wide range of treatment modalities (…) Patients in the study included approximately 80 percent of all admissions during a ten-month average period. Most nonparticipants dropped out of their programs within a few days of admission." (Bromet & Moos, 1977, p. 328)
Sharing/Openness (i.e., materials, data, code).	2	All items of the measures are reported: "*Background Data* Shortly after admission to a program, patients were given a Background Information Form, a structured, self- administered questionnaire covering a broad array of personal and alcohol-related information. Sociodemographic characteristics, as well as the following items and subscales, were used in the present paper" (Bromet & Moos, 1977, p. 328-329) "Marital and occupational resources are measured in terms of their presence or absence, and, if present, the quality of the social environment experienced by the patients. The social climate of the family was measured using the Family Environment Scale, which obtains the perceptions of all family members in the home (Moos, 1974a). Four areas of the family environment were related to outcome-cohesion, conflict, moral-religious emphasis, and control. Perceptions of the social climate of work settings were measured using the Work Environment Scale (Moos and Insel, 1974) and were analyzed separately for married and unmarried alcoholics. Based on the suggestions by

Kasl (1974), seven environmental areas were related to outcome-involvement, peer cohesion, staff support, autonomy, work pressure, clarity, and physical comfort." (Bromet & Moos, 1977, p. 328)

"Four outcome criteria were selected to provide data on the major dimensions of functioning: behavioral impairment, subjective rating of drinking problem, psycho- logical well-being, and social functioning.
Family environment. Patients who were living with their families were asked, along with the other family members (12 years of age or older) in the home, to fill out the Family Environment Scale (FES) composed of 90 true-false items that evaluate the social climate of all types of families. The FES is composed of 10 sub- scales which measure the interpersonal relationships among family members, the directions of personal growth which are emphasized in the family, and the basic organizational structure of the family. (…) Four subscales were the focus of the present analysis: (1) *Cohesion,* or the extent to which family members are concerned, helpful and supportive of each other. (2) *Conflict,* or the extent to which the open expression of anger and conflictual interactions are characteristic of the family. (3) *Moral-Religious Emphasis,* or the extent to which the family actively discusses and emphasizes ethical and religious is- sues and values. (4) *Control,* or the rigidity of family rules and procedures and the extent to which family members order each other around. (…)
Work Environment. Patients who were employed in nonsolitary occupations at the time of follow-up were asked to fill out a Work Environment Scale (WES). (…) Seven subscales were the focus of the present analysis: (1) *Involvement,* or the extent to which workers feel committed to their jobs. (2) *Peer Cohesion,* or the extent to which workers are friendly and supportive of each other. (3) *Staff Support,* or the extent to which supervisors are supportive of workers. (4) *Autonomy,* or the extent to which workers are encouraged to make their own decisions. (5) *Work Pressure,* or the extent to which workers experience deadlines and excessive pressure to work hard. (6) *Clarity,* or the extent to which workers know what to expect in their daily routines and the extent to which rules and policies are explicitly communicated. (7) *Physical Comfort,* or the extent to which the physical surroundings

		contribute to a pleasant work environment" (Bromet & Moos, 1977, p. 330)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	1	"*Analyses* The relationships between the availability of environmental resources at ad- mission (marital and employment status) and posttreatment performance were analyzed using One-Way Analyses of Variance. (The sample did not lend itself to a 3 x 2 design [employment stability x married/unmarried] because of an in- adequate number of patients in one cell [married-unemployed; n=2].) Relation-ships between perceptions of family and work environments and posttreatment adjustment were analyzed using product-moment correlations." (Bromet & Moos, 1977, p. 331)
Fallacious interpretation of (lack of) statistical significance.	2	"The significant differences (One-Way Analysis of Variance) on each of the criterion variables indicates that patients who were married or widowed at admission had more positive outcomes than patients in the remaining three groups. Two sub-analyses further supported this finding. One-Way Analyses of Variance were computed on the three unmarried categories, and no significant differences were found. T-tests comparing the married and widowed patients with the divorced, separated, and single patients as a group were all highly significant ($p < .001$). Thus, marital resources at admission were predictive of more favorable outcome." (Bromet & Moos, 1977, p. 331)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	The paper does not disclose enough statistics to calculate the single-article *p*-curve.

Number 3 of the Center 10

The number 3 of the center 10 is the first study reported in the paper *Self-Reported Attachment Patterns and Rorschach-Related Scores of Ego Boundary, Defensive Processes,*

and Thinking Disorders (Berant & Wald, 2009). Table C3 shows the score on each of the eight selected RDF for the number 3 paper.

Table C3

Coding paper nr. 3 of the center 10 (i.e., Berant & Wald, 2009)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	"However, the Berant et al. (2005) study focused solely on variables derived from Exner's (2005) Comprehensive System (CS). In this study, we focus on boundary issues, making use of both CS scores and additional scales. We examine the associations between the individual's attachment style and content-based dimensions of ego-boundary representations and defensive processes. We begin by briefly reviewing the basic principles of attachment theory and describing the associations between attachment styles and Rorschach CS scores. Next, we describe the concepts of ego boundary, defensive processes, and CS Rorschach thought disorder scores reflecting ego-boundary disturbance while hypothesizing their associations with attachment styles." (Berant & Wald, 2009, p. 365) Confirmatory: "To summarize, our hypotheses are that anxiously attached individuals, who wish to merge with relationship partners and are highly sensitive to their surroundings and to the threat of being abandoned, have a "thin skin" in their interactions with the world and will display Rorschach responses that reveal high penetration scores. In addition, anxiously attached people are passive in their coping and are easily given to the influence of external reality and will thus manifest thinking problems that characterize a tendency toward boundary blurring. We thus hypothesize that the higher the individuals' attachment anxiety, the more Incongruous and Fabulized Combinations (INCOM and FABCOM) will be displayed in their Rorschach protocols scored using Exner's (2005) CS system. Anxiously attached individuals are also expected to reveal higher projective identification scores on the Rorschach because they tend to "hold on" to the other, controlling him or her so as not to feel deserted. Regarding attachment avoidance, we hypothesize that avoidantly attached individuals,

who are reluctant to show weakness and neediness, will reveal higher devaluation scores on the Rorschach. By diminishing the value of the other, they remain strong and untouched. Avoidant individuals are also expected to reveal higher splitting defenses scores on the Rorschach because they attempt to keep their defensive, grandiose self-esteem from being "contaminated" by negative feelings and thoughts." (Berant & Wald, 2009, p. 367)

"Penetration scores have more mixed evidence concerning their validity (O'Neill, 2005). However, as an index that captures representations of permeable boundaries, we anticipated it would be correlated with anxious attachment, as persons with an anxious attachment orientation seek to merge with others as a means of coping with the experience of insecurity (Bartholomew & Horowitz, 1991). They also are less able to repress negative affect, have easier access to negative memories, and are more vulnerable to emotional spreading (Mikulincer & Orbach, 1995).
A Rorschach scoring system designed to evaluate the specific operations presumed to characterize the developmental level of defensive functioning was introduced by Lerner and Lerner (1980). Lerner and Lerner's conceptualization was based on Kernberg's (1975) theoretical model of defense that conceived of internal object relations as organized on the basis of specific defensive operations. In this study, we examined defenses that we assume are typically used by attachment anxious individuals (projective identification) and avoidant individuals (devaluation and splitting) to deal with distress and relationship dilemmas." (Berant & Wald, 2009, p. 366-367)

"As described previously, the associations found in Berant et al.'s (2005) study were based on Rorschach scores derived from Exner's (2005) CS. To widen the scope of this line of research and to understand further the dynamics of the activation of the attachment system, in this article, we focus on boundary issues. To do so, we make use of both content-based scales and alternative CS scores that can be theoretically related to attachment styles. In this article, we examine the associations between attachment-related anxiety and avoidance measures on one hand and these Rorschach markers of ego boundary and defensive processes on the other.

		Specifically, we focus on scales that measure body boundary representations (Barrier and Penetration scales; Fisher & Cleveland, 1958) and the use of defenses of projective identification, splitting, and devaluation (Lerner, 1998). In addition, we examine CS markers of thinking difficulties that are also thought to reflect boundary blurring by psychoanalytically oriented Rorschach scholars (Blatt & Ritzler, 1974; Lerner, 1985; Schafer, 1954). Specifically, we focus on the CS scores of Incongruous Combinations (INCOM) and Fabulized Combinations (FABCOM)." (Berant & Wald, 2009, p. 366) "In this study, we addressed associations between self-reported attachment scales (anxiety and avoidance) and Rorschach (1921/1942) indexes indicating ego-boundary perception (barrier and penetration), use of projective identification, devaluation and splitting defenses, and Comprehensive System (Exner, 2005) scores that represent boundary blurring (incongruous and fabulized combinations)." (Berant & Wald, 2009, p. 365)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	"A nonclinical sample of 89 citizens of Israel (women and men) ranging in age from 19 to 57 years (median=23 years) participated in this study without monetary reward." (Berant & Wald, 2009, p. 367) "This sample was comprised of the 72 participants who had participated in the original Berant et al. (2005) study plus an additional 17 student participants with similar background and demographic characteristics as the first group. Whereas the original 72 participants also served as part of a normative sample of 150 Israelis in an additional study (Berant, 2007), the 17 new

		participants did not participate in any additional study." (Berant & Wald, 2009, p. 368) "In this study, we extended the sample and findings described by Berant, Mikulincer, Shaver, and Segal (2005) using a nonclinical sample of 89 Israeli adults." (Berant & Wald, 2009, p. 365)
Sharing/Openness (i.e., materials, data, code).	3	"*Materials and Procedure* We conducted the study in two sessions. In the first session, participants were tested in small groups and were asked to complete Mikulincer, Florian, and Tolmacz's (1990) 10-item Hebrew language version of the Hazan and Shaver's (1987) scale measuring attachment anxiety and avoidance in close relationships. This scale includes five items tapping avoidant attachment (e.g., "I am somewhat uncomfortable being close to others") and five items tapping anxious attachment (e.g., "I often worry that my partner doesn't love me"; "I find that others are reluctant to get as close as I would like"). Items were constructed based on Hazan and Shaver's (1987) prototypical descriptions of attachment styles and were highly similar to the English-language ECR scales (Brennan et al., 1998). Participants were asked to think about their close relationships without focusing on a specific partner and to rate the extent to which each item described them in these relationships on a 7-point scale ranging from 1 (*not at all*) to 7 (*very much*). (…) The second session was conducted 1 month later by different research assistants than those who had administered the attachment scales. Participants were seen individually and were asked to complete the Rorschach Inkblot Test. The examiners (3 senior graduate students from the clinical psychology program at Bar-Ilan University) had taken three basic and advanced courses in personality assessment and were familiar with administering, coding, and analyzing the Rorschach according to Exner's (1995, 2001) CS system. They were unaware of the participants' attachment orientation and were not acquainted with any participant. Before the study began, the examiners underwent an additional 6 hr of specific training with E. Berant to ensure standardization of administration, coding, and scoring" (Berant & Wald, 2009, p. 368)

Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	1	"As can be seen in Table 1, the anxiety attachment and avoidance attachment scores were distributed normally in this sample, with means of 3.28 and 3.34 and standard deviations of 1.12 and 1.05, respectively." (Berant & Wald, 2009, p. 368) "To test our hypotheses about the associations between self-reported attachment dimensions and content-based Rorschach scales, we conducted a two-step procedure. In the first step, we computed Pearson correlations between self-reported attachment anxiety and avoidance with the following Rorschach scales: Barrier, Penetration, Projective Identification, Splitting, Devaluation, Level 1 Incongruous Combinations, and Level 1 Fabulized Combinations. In the second step, we conducted a multiple regression analysis examining the unique contributions of attachment scores to each of the Rorschach indexes. In this regression, we introduced attachment anxiety and avoidance simultaneously as predictors, and we introduced each Rorschach score as the predicted variable. In this way we examined the extent to which attachment anxiety and avoidance predicted each of the Rorschach scores." (Berant & Wald, 2009, p. 369)
Fallacious interpretation of (lack of) statistical significance.	2	"The multiple regression results point to the contribution of attachment anxiety or avoidance to the explained variability. presents [sic] the relevant Pearson correlations and standardized regression coefficients (beta) of the relevant analyses. As can be seen in Table 2, attachment anxiety is positively associated with and made a unique contribution to the use of Projective Identification, Penetration, INC, and FAB. That is, the higher the level of attachment anxiety the more use there is of projective identification, the more contents of penetration, and the more CS scores suggesting boundary blurring. Attachment avoidance approximates significant positive association with the use of devaluation, and it approaches a significant unique contribution to the use of devaluation. That is, the higher the level of attachment avoidance, the more there is a tendency to use devaluation. Attachment dimensions (anxiety

		or avoidance) did not have significant associations with Barrier contents or with the defense of splitting." (Berant & Wald, 2009, p. 369-370)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	The paper does not disclose enough statistics to calculate the single-article *p*-curve. Table 2: "Pearson correlations and standardized regression coefficients (betas) for predicting Rorschach scores from both attachment anxiety and attachment avoidance." (Berant & Wald, 2009, p. 370)

Number 4 of the Center 10

The number 4 of the center 10 is the second study reported in the paper the second study reported in the paper *A license to speak up: Outgroup minorities and opinion expression* (Morrison, 2011). Table C4 shows the score on each of the eight selected RDF for the number 4 paper.

Table C4

Coding paper nr. 4 of the center 10 (i.e., Morrison, 2011)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "Across all studies, I hypothesized that outgroup minorities would be more likely to express disagreement with other group members when their social category granted them psychological standing on the issue ((…) men in Study 2 (…)) than when it did not ((…) women in Study 2 (…)." (Morrison, 2011, p. 758) "It was predicted that male participants who imagined disagreeing with women on gays in the military would report being more likely to express their opinions than female participants who imagined disagreeing with men, as the former should have more psychological standing the issue." (Morrison, 2011, p. 760) "In the present studies, I test the hypothesis that outgroup minorities are especially inclined to speak up when they believe, by virtue of their social category membership, that they have the psychological standing to do so (Miller, 1999; Miller, Effron, & Zak, 2009). Examining this

		potential boundary condition will not only help determine when outgroup minorities are most likely to make their opinions heard, but will also shed light on how to encourage different groups of people to voice their opinions on controversial issues." (Morrison, 2011, p. 757) "The present studies were conducted to determine when outgroup minority status – that is, being both an outgroup member and an opinion minority – would and would not increase opinion expression on controversial issues. Specifically, the objective was to test the role of psychological standing, as determined by social category membership. Several different social categories and types of issues were examined: (…) gender and a male-relevant issue in Study 2" (Morrison, 2011, p. 757)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"One participant who did not complete the opinion expression measure was omitted, leaving 151 individuals in the final sample." (Morrison, 2011, p. 759)
Sample size (predetermined or not).	3	"One hundred and fifty-two individuals (56 men, 96 women; M_{age} = 35.7, SD = 11.7), all U.S. citizens, were recruited for this study through the same email database as in Study 1. Participants were randomly assigned to either the outgroup minority condition or the ingroup minority condition. In exchange for participation, they were entered into a drawing to win one of several $25 gift certificates to a major online retailer." (Morrison, 2011, p. 759)
Sharing/Openness (i.e., materials, data, code).	2	"In Study 2, male and female participants read a hypothetical scenario in which a group of either men or women disagreed with their opinion on gays in the military. Afterward, they reported how likely they would be to voice their opinions in this situation." (Morrison, 2011, p. 758) "All experimental materials were presented online. Participants first read a short scenario similar to that used in Study 1. However, the scenario was re-worded so that it pertained to whether gays should be allowed to serve in the military, rather than to

		affirmative action. The scenario instructed participants to imagine that they were standing in a group at a party where they did not know many people, that the others in this group had begun discussing whether or not gays should be allowed to openly serve in the military, and that someone in the group asked them for their opinion. This issue was chosen because the military is a predominantly male institution, and controversies about gays in the military are more often discussed in terms of gay men than lesbians (see Gonzenbach, King, & Jablonski, 1999). Male participants should thus have more psychological standing on the issue than female participants. Because there were no significant effects of social category status or group membership (i.e., race) within the opinion majority condition in Study 1, all participants in Study 2 read a scenario in which they were opinion minorities. In other words, all participants read that the other group members held a different opinion than they did on the issue of gays in the military. (…) Next, participants indicated how likely they would be to (1) express their true opinion in the situation, (2) try to change the topic, and (3) walk away from the group (1 = *not at all likely*, 5 = *very likely*) in response to being asked their opinion. After reverse-scoring the last two items, participants' responses were averaged to form an index of minority opinion expression (α=.60). A principal components analysis revealed that the three items loaded onto a single factor, which explained 55.81% of the total variance (eigenvalue=1.67). At the end of the experiment, participants completed a demographic questionnaire in which they reported their age, gender, and attitude toward allowing gays to openly serve in the military (1 = *extremely opposed*, 9 = *extremely in favor*; *M* = 6.74, *SD* = 2.09). As in Study 1, participant attitude was included as a factor in the analysis. Additionally, to check whether men were perceived as having greater psychological standing on this issue than women, participants indicated whether they believed the issue of gays in the military was more relevant to men or women (1 = *much more relevant to men*, 7 = *much more relevant to women*).” (Morrison, 2011, p. 759-760)
Using covariates and reporting the results	0	No covariates.

with and without the covariates.		
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"participants' scores on the opinion expression measure were submitted to a social category condition (0 = ingroup minority, 1 = outgroup minority) × participant gender (0 = male, 1 = female) × participant attitude (mean-centered continuous variable) multiple regression analysis." (Morrison, 2011, p. 760)
Fallacious interpretation of (lack of) statistical significance.	2	"The results of Study 2 conceptually replicated those of Study 1, using a different social category (gender) and issue (gays in the military). Specifically, the results demonstrated that men who disagreed with a group of women on gays in the military reported being more likely to express their opinions than did women who disagreed with a group of men. That is, outgroup minorities were more emboldened to express their opinions to the extent that they had more psychological standing than other group members (as was the case for male participants in a group of women). Both Studies 1 and 2 have provided evidence that outgroup minorities are more willing to express their opinions when they have more psychological standing than other group members than when they have less standing." (Morrison, 2011, p. 760)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	2	"*Manipulation check* First, to ascertain that the issue of gays in the military was seen as more relevant to men than women, a one-sample *t* test was performed on this item. The mean rating was significantly below the scale midpoint of 4 (M = 3.49, SD = 1.12), $t(150) = -5.57$, $p < .001$, suggesting that men were thought to have more standing on the issue. Furthermore, both male participants (n=57, M=3.47, SD=1.09), $t(56) = -3.65$, $p < .001$, and female participants (n=94, M=3.50, SD=1.15), $t(93) = -4.21$, $p < .001$, perceived men as having more standing than women. *Opinion expression* The predicted interaction between social category condition and participant gender was significant (β=.42), $t(144)=2.58$, p=.01 (see Fig. 2). As hypothesized, male participants who imagined disagreeing with women (i.e., male outgroup minorities) reported a marginally greater likelihood of expressing their opinions than did female participants who imagined disagreeing with men

(i.e., female outgroup minorities) (β=−.23), $t(144)$ = -1.92, p=.057. Conversely, male ingroup minorities (i.e., who imagined disagreeing with a same-gender group) reported being marginally less likely to express their opinions than did female ingroup minorities (β=.18), $t(144)$=1.65, p=.10. No other lower-order effects were significant, nor was the three-way interaction between social category condition, participant gender, and participant attitude (β=−.03), $t(144) < 1$, *ns*." (Morrison, 2011, p. 760)

In Figure C3, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $t(144) = 2.58$; $t(144) = -1.92$; $t(144) = 1.65$.

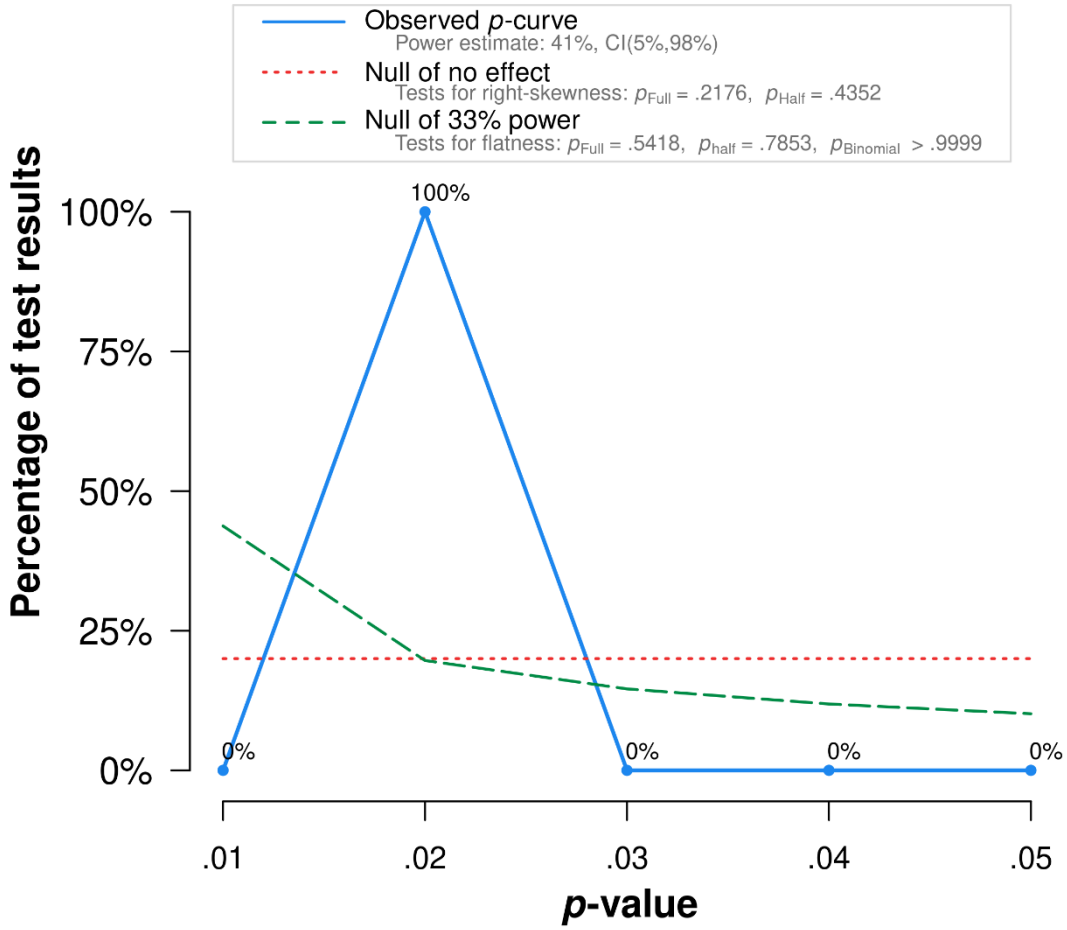


Figure C3. The single-article *p*-curve for the number 4 of the center 10 (i.e., Morrison, 2011).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C4, not only is the half *p*-curve test ($p = .4352$) not significantly right-skewed ($p < .05$), but also both the half ($p = .4352$) and full test ($p = .2176$) are not significantly right-skewed ($p < .1$), which implies that the study does not contain evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C4, the 33% power test is $p = .5418$ for the full *p*-curve, for the half p-curve is $p = .7853$, and for the binomial 33% power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.5	*Z*=-0.78, *p*=.2176	*Z*=-0.16, *p*=.4352
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*>.9999	*Z*=0.11, *p*=.5418	*Z*=0.79, *p*=.7853
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 41% 90% Confidence interval: (5% , 98%)		

Figure C4. Additional statistics for the single-article *p*-curve for the number 4 of the center 10 (i.e., Morrison, 2011).

Number 5 of the Center 10

The number 5 of the center 10 is the first study reported in the paper *When Stigma Confronts Stigma: Some Conditions Enhancing a Victim's Tolerance of Other Victims* (Galanis & Jones, 1986). Table C5 shows the score on each of the eight selected RDF for the number 5 paper.

Table C5

Coding paper nr. 5 of the center 10 (i.e., Galanis & Jones, 1986)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*

Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: Having thus briefly sketched the elements of our experimental design, let us summarize by presenting the major hypothesis of the study. In statistical terms, we predicted a significant higher-order interaction effect involving (1) subject race, (2) victimization salience, and (3) mental patient label. We predicted that blacks would be more favorable than whites toward an ex-mental patient, but only to the extent that their own potential victimization as a black person was made salient. Because the victimization message also suggested that deviants can be dangerous, the message was predicted to make whites less sympathetic than blacks (and than whites in the absence of the message) toward an ex-mental patient. In addition, it was predicted that blacks would be less sympathetic than whites in the absence of victimization salience, following the assumption of "displaced retribution" mentioned above. (Galanis & Jones, 1986, p. 171) The present experiment was designed to explore the conditions that might enhance the tendency of markables to react with sympathy to other markables. (Galanis & Jones, 1986, p. 170) *Do blacks differ from whites in their ascription of negative stereotypes to former mental patients? Because blacks are aware that they themselves are often the victims of negative stereotyping, they might show greater reluctance to apply negative stereotypes to others marked by society as "deviant." Black and white* (Galanis & Jones, 1986, p. 169) The present investigation concerns the impact of one form of markability on reactions to a target person who is also markable, but in a very different way. The question posed is whether one's own vulnerability to the stigmatizing process makes one more, or less, sympathetic to other markables. It is not (Galanis & Jones, 1986, p. 169)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	**Subjects** Subjects were 40 white and 40 black undergraduates at Princeton University. The (black) experimenter approached them individually in the dining halls and student activity centers of the residential colleges and asked them to participate in a study on "civil commitment." Those who agreed to participate were handed one of four booklets containing all instructions and experimental materials, including a "face sheet" on which subjects identified their sex, age, race, and expected year of graduation. Booklets were randomly ordered with the constraint that equal numbers of blacks and whites appeared in each condition. Except for the race variable, the experimenter was blind to experimental condition. (Galanis & Jones, 1986, p. 172) Black and white undergraduates were recruited for an experiment under the assumption that black students in a predominantly white student body are more aware of their markability than are majority white students. Blacks, by and (Galanis & Jones, 1986, p. 170)

Sharing/Openness (i.e., materials, data, code).	2	The introductory section of the questionnaire described the experimenter's interest in the criteria for commitment of the mentally ill to institutional care. It noted that there are two kinds of commitment, criminal and civil. A brief discussion of criminal commitment noted some of the reasons for commitment. Half of the booklets at this point contained a case example to indicate the complexity of criminal commitment decisions. This was the *salience* manipulation in that the case was that of a black Harlem teenager who, after a history of poverty and family disintegration leading to delinquency and burglary, shot and killed a jogger for no apparent reason. This supposedly illustrative case vignette ended with the sentence: "Leroy's lawyer entered a plea of insanity on the basis that he did not know right from wrong and that, like many blacks, Leroy was a victim of society." The instructions then turned to the "main focus of the study," civil commitment, noting recent controversy over whether or not lay persons are sufficiently qualified to make appropriate commitment decisions. An actual case was to be presented and the subject's decision with regard to commitment would be compared with that of the commitment board. The case itself (fabricated for the experiment) was that of a 30-year-old construction worker whose behavior had undergone significant changes in the past 6 months, including weight loss, fatigue, depression, and weekend binge drinking. The case report noted that he often became loud and abusive during his binges and stated that he was referred for commitment by a friend who was concerned by his reports of auditory and visual hallucinations. A final sentence cautioned that "mental institutions have had little success in treating problems of this kind." The race of the worker was not mentioned. Subjects in the label condition were also informed that the target person had previously spent 2 years committed to a mental institution, after which he adjusted reasonably well until the recent episodes began. Subjects in the no-label condition were instead told, "During adolescence, he experienced considerable emotional problems." (Galanis & Jones, 1986, p. 172) There were five interrelated dependent measures: (1) a rating of whether the target person should be committed, (2) a rating of illness extremity, (3) a rating of the target person's "chances of leading a normal life," (4) a rating of the degree of danger that target person posed to the general public, and (5) how willing the subject would be to have the target person as a neighbor or job partner. When they completed these ratings and returned the questionnaire, subjects were informed concerning the actual purpose of the study. (Galanis & Jones, 1986, p. 173) White and black subjects were asked to make several evaluative judgments about a target person experiencing serious adjustment difficulties. For half the subjects this was preceded by information about a black teenage defendant designed to make salient the possibilities of victimization. Cross-cutting this manipulation, information about the target person either did or did not include a reference to the fact that he had previously spent 2 years in a mental institution. (Galanis & Jones, 1986, p. 171)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	1	Thus the experiment involved a 2 (subject race) × 2 (victimization salience) × 2 (mental illness label) between-subjects factorial design. (Galanis & Jones, 1986, p. 171) should probably be committed. Because the dependent measures are theoretically interrelated, a multivariate analysis of variance was performed to examine the overall effect of the experimental manipulations of salience and labeling for the two races. The results of this analysis are presented in Figure 1 and Table 1, (Galanis & Jones, 1986, p. 173)

Fallacious interpretation of (lack of) statistical significance.	3	Reporting about the magnitude of the interaction without reporting effect sizes: **Here we may note a number of significant main effects and interactions. The significant race × salience × label interaction confirms the major hypothesis of the study. When their own potential or actual victimization is made salient by a vivid example of a black delinquent with a history of neglect and maltreatment and when the target person is given a stigmatizing label, black subjects show a high level of tolerance. In contrast, the combination of salience and labeling makes white subjects *less* tolerant than in the absence of these conditions. The magnitude of the simple race by salience interaction suggests that there is a general tendency for whites and blacks to respond very differently to the priming description of a black delinquent on trial for undisputed homicide. Figure 1 shows clearly that the mental illness label augments this general tendency. Indeed, it does so strongly enough to produce a qualifying triple interaction.** (Galanis & Jones, 1986, p. 173)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	2	TABLE 1 Responses to Target Person: Analysis Summaries *Univariate F-Ratios (df 1, 72)*[a] / *Mv*[b] *Source* — *(1)* — *(2)* — *(3)* — *(4)* — *(5)* — *Mv*[b] Race — 3.26 — 2.83 — 1.50 — .02 — .09 — .88 Salience — 4.03* — .11 — 2.93 — 9.17* — 6.00* — 4.71* Label — 21.30* — 1.81 — 4.84* — .52 — .09 — 4.33* R × S — 1.97 — 1.81 — 8.61* — 45.91* — 37.50* — 18.80* R × L — 1.45 — .11 — 2.15 — 4.68* — 2.34 — 2.05 S × L — 4.87* — 2.83 — 3.83 — 3.51 — 1.50 — 5.01* R × S × L — 4.03* — 1.81 — 4.84* — 9.17* — 3.38 — 3.55* a. (1) Should TP be committed? (2) How ill is TP? (3) What are TP's chances of ever leading normal life? (4) How dangerous is TP to general public? (5) How willing would you be to have TP as neighbor or job partner? b. Multivariate analysis F-ratios, df = 5, 68. *p < .05. **highly significant for perceived dangerousness, F(1, 72) = 21.99, p < .001, and desire to avoid, F(1, 72) = 16.92, p < .001. Thus, when there is no prior exposure to the case of the black delinquent, the black subjects are *more* fearful and avoidant of the disturbed and maladjusted target person than the white subjects.** (Galanis & Jones, 1986, p. 173)

Because attenuation of the attenuated interaction is predicted in the 2 x 2 x 2 design, only the three-way interaction is selected as input for the *p*-curve (Simonsohn et al., 2015). In Figure C5, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(5, 68) = 3.55$.

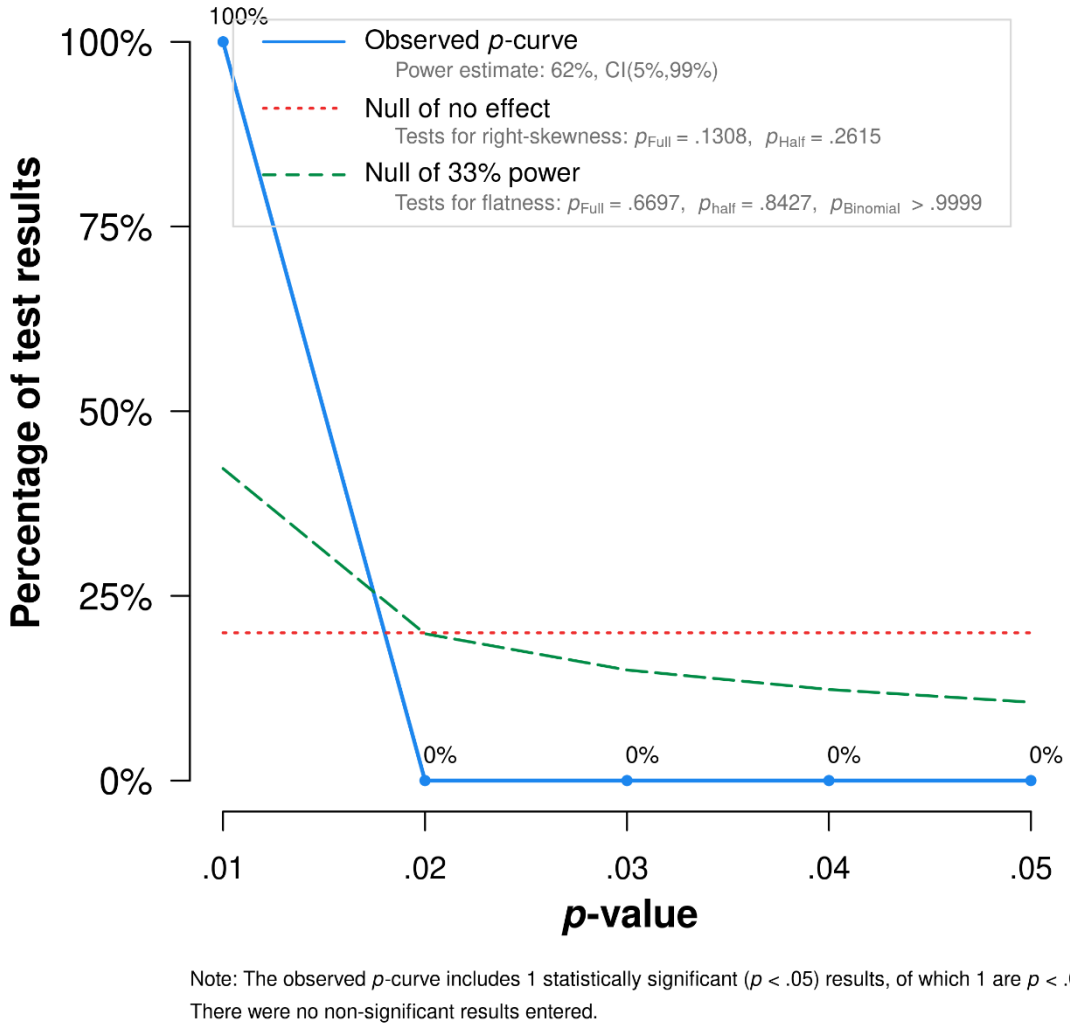


Figure C5. The single-article *p*-curve for the number 5 of the center 10 (i.e., Galanis & Jones, 1986).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C6, not only is the half *p*-curve test ($p = .2615$) not significantly right-skewed ($p < .05$), but also both the half ($p = .2615$) and full test ($p = .1308$) are significantly right-skewed ($p < .1$), which implies that the study does not indicate evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C6, the 33% power test is $p = .6697$ for the full *p*-curve, for the half p-curve is $p = .8427$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.5	Z=-1.12, p=.1308	Z=-0.64, p=.2615
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p>.9999	Z=0.44, p=.6697	Z=1.01, p=.8427
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 62% 90% Confidence interval: (5% , 99%)		

Figure C6. Additional statistics for the single-article *p*-curve for the number 5 of the center 10 (i.e., Galanis & Jones, 1986).

Number 6 of the Center 10

The number 6 of the center 10 is the first study reported in the paper *Consideration of future consequences scale: Confirmatory Factor Analysis* (Hevey et al., 2010). Table C6 shows the score on each of the eight selected RDF for the number 6 paper.

Table C6

Coding paper nr. 6 of the center 10 (i.e., Hevey et al., 2010)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "Individual differences in the Consideration of Future Consequences (CFC) are typically assessed using the 12-item scale developed by Strathman, Gleicher, Boninger, and Edwards (1994). However, in contrast to the unidimensional model proposed by the scale developers, recent factor analyses have produced two-dimensional models of the scale. Confirmatory factor analyses were used in this study to evaluate different 1- and 2-factor models based on data provided by 590 (236 males, 354 females) young adult members of the general public." (Hevey et al., 2010, p. 654) "the hypothesized factor models" (Hevey et al., 2010, p. 655) "Despite the widespread use of the CFC scale, it has received relatively little psychometric evaluation. Given the discrepancies between the findings in

		relation to the underlying factor structure of the CFC and the potential for method effects, the present study tested a number of competing models using Confirmatory Factor Analysis. We examined the fit of the original unidimensional 12- item model, a unidimensional 8-item model (based on Petrocelli (2003)), and unidimensional model incorporating correlated errors to reflect measurement artifacts. In addition, a two-dimensional model based on Petrocelli (2003), and two-dimensional model based on Joireman et al. (2008) were examined. It should be noted that these models differ only in relation to 1-item (item 2, a positively coded item), which is included with all the reverse coded items by Petrocelli, whereas it is included with all the positively coded items by Joireman et al." (Hevey et al., 2010, p. 655)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	2	"*Participants* Convenience sampling was used to recruit 590 (236 males, 354 females) members of the general public in Ireland, aged between 16 and 26 years (M = 20.4 years, SD = 3.1). Approximately 2/3 of the participants were currently higher education students. Recruitment took place between December 2007 and January 2008 in various locations around Ireland (e.g., schools, sports clubs, colleges and train stations). A script was used to ensure that all participants were approached in the same way." (Hevey et al., 2010, p. 655)
Sharing/Openness (i.e., materials, data, code).	2	All 12 items are shared in Table 2 (Hevey et al., 2010, p. 656). "*Measures* The first section of the questionnaire recorded demographic details such as age, gender, occupation and education. Section 2 consisted of the 12-item CFC scale, which comprises both positively (5 items) and negatively (7 items)

		phrased items; respondents indicate the extent to which each statement is characteristic of their behaviour, with response options ranging from "*Extremely uncharacteristic*" (1) to "*Extremely characteristic*" (5). The scale ranges from 12 to 60; a high CFC score indicates a high degree of importance being placed on the future consequences of behaviour, whereas low CFC score indicates greater importance being placed on the more immediate consequences of behaviour. Cronbach's alpha was .82 for the scale in the present sample." (Hevey et al., 2010, p. 655)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"*Data analysis* Confirmatory Factor Analysis was conducted using AMOS 7 (SPSS Inc., Chicago, Illinois 60606, US). The chi-square index is inadequate as a stand alone fit index because of its sensitivity to both small and large sample sizes (Bentler & Bonett, 1980) and therefore a variety of fit indices were used to evaluate the hypothesized factor models. The standardised root mean square residual (SRMR), which quantifies the mean absolute value of the correlation residuals, is also reported; lower values indicate better model fit, with values below .05 indicating good model fit. Furthermore, model fit was examined in relation to the goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), normed fit index (NFI) and comparative fit index (CFI), which all approach 1 for a perfect model fit. Values around .95 or higher are typically taken to indicate good fit of the model to the data (Hu & Bentler, 1999). The root mean square error of approximation (RMSEA) is a parsimony adjusted index that corrects for model complexity and should be lower than .05 to indicate a close approximate fit (Hu & Bentler, 1999)." (Hevey et al., 2010, p. 655)
Fallacious interpretation of (lack of) statistical significance.	2	"The data support the CFC as being unidimensional, but with method effects influencing responses to positive and negative items." (Hevey et al., 2010, p. 655) "The results of the present study suggest that the separation between two empirically identified factors may reflect method effects associated with the use of item wording. Method effects are

		systematic variance that is attributable to the measurement method rather than to the constructs the measures represent (Podsakoff et al., 2003). A typical instance of method effects in questionnaire data are response styles which can decrease the correlations of positively and negatively worded items, thus leading to the lack of fit of a unidimensional model, if response styles affect positively and negatively worded items differently. In accordance with the method effects explanation, lack of fit of a unidimensional model is a common finding for questionnaires with positively and negatively phrased items. The evidence presented here indicated that the CFC scale items reflect one substantively meaningful construct and substantively irrelevant method effects." (Hevey et al., 2010, p. 656) "The general CFC model with method effects was more parsimonious than positing two discrete factors." (Hevey et al., 2010, p. 657)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	**Table 1** Goodness-of-fit indices for CFC models in present data and previous reports. (Hevey et al., 2010, p. 655)

Model	χ^2 (df)	SRMR
1 Factor: 12 items	338.94 (54)	.061
1 Factor: 8 items	86.38 (20)	.041
Petrocelli (2003) data	106.48 (20)	–
1 Factor with correlated errors	61.41 (24)	.025
2 Factors: Petrocelli (2003)	294.99 (53)	.059
Petrocelli (2003) data	232.07 (53)	–
2 Factors: Joireman et al. (2008)	209.72 (53)	.047
Joireman et al. (2008) data	294 (53)	.042

In Figure C7, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $\chi^2(54) = 338.94$; $\chi^2(20) = 86.38$; $\chi^2(24) = 61.41$.

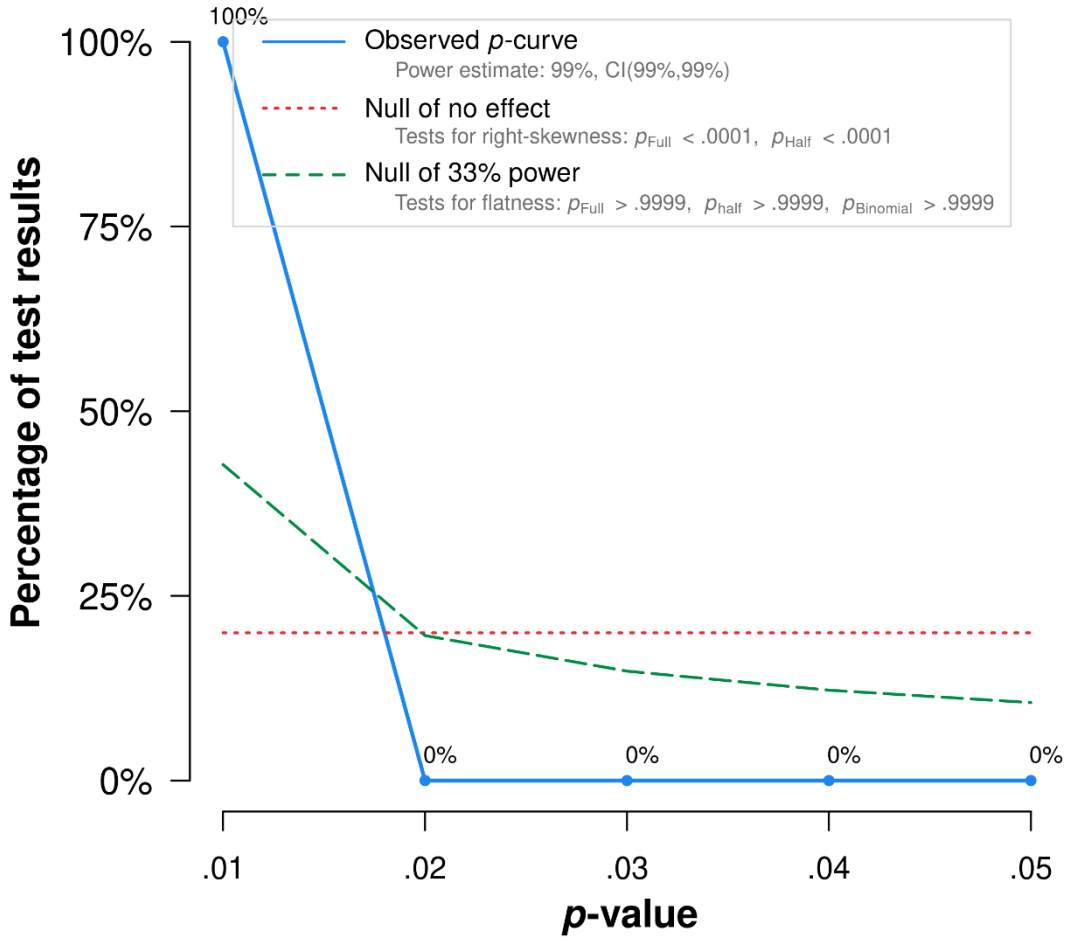


Figure C7. The single-article *p*-curve for the number 6 of the center 10 (i.e., Hevey et al., 2010).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C8, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C8, the 33% power test is $p > .9999$ for the full *p*-curve, for the half p-curve, and for the binomial 33% power test; "so

p-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.125	Z=-9.58, p<.0001	Z=-9.34, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p>.9999	Z=8.52, p>.9999	Z=8.77, p>.9999
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure C8. Additional statistics for the single-article *p*-curve for the number 6 of the center 10 (i.e., Hevey et al., 2010).

Number 7 of the Center 10

The number 7 of the center 10 is the first study reported in the paper *Social Identity, Modern Sexism, and Perceptions of Personal and Group Discrimination by Women and Men* (Cameron, 2001). Table C7 shows the score on each of the eight selected RDF for the number 7 paper.

Table C7

Coding paper nr. 7 of the center 10 (i.e., Cameron, 2001)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "In summary, the primary aim of this study was to determine the extent to which women's and men's perceptions of personal and group discrimination are jointly predicted by facets of gender-derived social identity (ingroup ties, centrality, ingroup affect) and modern sexism. It was expected that several factors would qualify the relationship between social identity and perceptions of discrimination. First, it was hypothesized that perceptions of discrimination would be positively related to the centrality of gender-category membership and perceptions of ingroup ties, and, consistent with the possibility of a motivational basis for claims of discrimination, negatively related to ingroup-derived affect. Second, a Gender x Neosexism effect was anticipated, with higher levels of personal and group discrimination

perceived by men who endorse modern sexist beliefs and by women who reject those beliefs. A third class of expectations was that this pattern would be manifest particularly for individuals for whom gender-category membership entails psychological centrality and strong ingroup ties, but also for individuals who evaluate that category relatively negatively (i.e., three-way interactions involving social identity, neosexism, and gender). By implication, social identity was expected to predict perceptions of discrimination—in a positive direction for centrality and ingroup ties and in a negative direction for ingroup affect—particularly for men who endorse modern sexist beliefs, and, conversely, for women with low levels of modern sexism. Finally, it was expected that social identification would generally be more robustly related to perceptions of group, rather than personal, discrimination (Postmes et al., 1999)." (Cameron, 2001, p. 750-751)

"There are, however, a number of factors that might qualify the relationship between social identity and perceptions of discrimination. Three such factors, which inform the hypotheses of this study, will be considered in turn: (a) the specific contribution of gender-category membership to identity (ingroup ties, centrality, or ingroup affect), (b) the extent to which modern sexist beliefs are endorsed, and (c) the level at which discrimination is perceived (personal or group)." (Cameron, 2001, p. 747)

"It was predicted, then, that perceptions of discrimination would be predicted by the centrality and ingroup ties aspects of social identity in a positive direction, but by ingroup affect in a negative direction." (Cameron, 2001, p. 748)

"It was expected, then, that higher levels of discrimination would be perceived by men who endorse modern sexist beliefs (i.e., those who are presumably motivated to preserve their dominant status) and by women who reject modern sexist beliefs (i.e., those who perceive gender-related inequality to be illegitimate). Moreover, it was hypothesized that this joint effect of gender and modern sexism would be particularly apparent for individuals who are also inclined by their gender-derived social identification to attend to and report gender-related discrimination." (Cameron, 2001, p. 748)

		"Thus, there are theoretical reasons to expect that social identity will be associated with intergroup comparisons (including those that highlight group discrimination) to a greater extent than with interpersonal comparisons (i.e., those that would be relevant to judgments of personal discrimination; Postmes, Branscombe, Spears, & Young, 1999)." (Cameron, 2001, p. 750) "Thus, although the primary goal of this study was to investigate the extent to which social identity and modern sexism predict perceived personal and group-level discrimination as independent criterion variables, a secondary aim was to provide an insight into the conceptual and methodological utility of the personal/group discrimination discrepancy." (Cameron, 2001, p. 750) "In this paper, perceptions of discrimination are examined with a focus on two aspects of what Deaux and LaFrance (1998) have referred to as the gender-related belief system: (a) social identification as women or men; that is, the strength and quality of psychological investment in the group; and (b) modern sexism; that is, beliefs regarding women in contemporary Western society. The central issue of interest is the extent to which perceptions of personal and group discrimination are predicted by the joint effects of social identity and modern sexism." (Cameron, 2001, p. 744) "Perceptions of gender-related discrimination against the self and group were examined in women and men, with a focus on the predictive utility of modern sexism and 3 dimensions of social identification (ingroup ties, centrality, and ingroup affect)." (Cameron, 2001, p. 743)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.

Sample size (predetermined or not).	3	"The sample comprised 321 undergraduates (206 women and 115 men; mean age = 20.05 years) at the University of Queensland. The majority (77.9%) identified themselves as White, and 9.7% were Asian. Participants signed up for a study on "social–psychological attitudes," and received course credit for completing questionnaires on two occasions separated by 1 week. Questionnaires were completed in mixed-sex groups of approximately 10–15 people. The items comprising the measures described below were embedded in random order in the first questionnaire." (Cameron, 2001, p. 751) "Questionnaires were completed by 321 undergraduates (206 women and 115 men), of whom 78% self-identified as White and 10% as Asian." (Cameron, 2001, p. 743)
Sharing/Openness (i.e., materials, data, code).	2	"Response options for all items ranged from 1 (*strongly disagree*) to 6 (*strongly agree*)." (Cameron, 2001, p. 751) Materials: "**Measures** *Perceived Personal Discrimination* Three items were used to assess perceptions of personal gender discrimination: "I have personally been discriminated against because I am a (wo)man," "I have personally been a victim of sexual discrimination," and "I consider myself a person who has been deprived of opportunities that are available to others because of my gender." The first of these is similar to the often single-item measures typically used in research on the personal/group discrimination discrepancy (e.g., Taylor et al., 1990), whereas the latter two were taken from Kobrynowicz and Branscombe (1997). Preliminary item and reliability analyses indicated that deleting the third item improved the internal consistency (Cronbach's alpha) of the measure from .70 to .83. For this reason, a composite score was computed as the mean of the first two items. *Perceived Group Discrimination* The perception of discrimination directed at one's gender group was assessed by the following three items (α = .68): "(Wo)men in Australia are, as a group, discriminated against," "(Wo)men in Australia have been systematically prevented from attaining their full potential" (Kobrynowicz & Branscombe, 1997), and "I do not believe that (wo)men today suffer from the effects of discrimination on the basis of sex" (Cameron &

Lalonde, 2001).

Social Identification

Gender-derived social identity was operationalized in terms of a three-factor model (Cameron, 2000) reflecting the following components: (a) ingroup ties (e.g., "I have a lot in common with other women"; α = .80), (b) centrality (e.g., "I often think about the fact that I am a man"; α = .72), and (c) ingroup affect (e.g., "In general, I'm glad to be a woman"; α = .78). Each subscale comprised four items, two of which were negatively phrased; these were recoded so that higher scores indicate greater identification (i.e., stronger ties, greater centrality, and more positive affect).

Modern Sexism

Modern sexist beliefs were assessed using the neosexism scale designed by Tougas et al. (1995); they define neosexism as "a manifestation of a conflict between egalitarian values and residual negative feelings toward women" (p. 843). Tougas et al. (1995) have demonstrated the discriminant validity of the neosexism scale vis-à-vis "old-fashioned" sexism with respect to the prediction of attitudes toward affirmative action. The neosexism scale also compares favourably with alternative measures (Campbell, Schellenberg, & Senn, 1997). In the present sample, Cronbach's α = .80. Higher scores indicate a relatively greater endorsement of modern-sexist beliefs.

Sex-Role Ideology

The Attitudes Toward Women Scale assesses beliefs regarding "the rights, roles, and privileges women ought to have or be permitted" (Spence & Helmreich, 1978, p. 39); as such, it can be considered one operationalization of "old-fashioned" sexism (Deaux & LaFrance, 1998). Responses to the 15-item version were averaged such that higher scores reflect more nontraditional (egalitarian) attitudes (α = .83).

Self-Esteem

The 10-item Rosenberg (1965) Self-Esteem Scale is a frequently used and well-validated measure of global, personal self-evaluation. Greater scores indicate more positive self-esteem (α = .85).

Social Desirability

Responses to the 33 items of the Marlowe–Crowne Social Desirability Scale (Crowne & Marlowe, 1960) were averaged so that greater scores indicate a tendency to respond in a relatively favourable manner (α = .81)." (Cameron, 2001, p. 751-753)

Using covariates and reporting the results with and without the covariates.	3	"Three additional variables were included in the analyses to control for individual differences on relevant dimensions: scores on the Attitudes Toward Women Scale (Spence & Helmreich, 1978), global self-esteem, and social desirability. The first of these was included primarily to control for "old-fashioned" sex-role beliefs (see Deaux & LaFrance, 1998; Swim et al., 1995). Global self-esteem is relevant to individual-level motivational explanations for claims of discrimination, particularly for advantaged group members (Kobrynowicz & Branscombe, 1997), and provides a useful point of comparison (as well as a relevant control variable) for the positivity of group-level self-evaluation. Finally, social desirability concerns were accounted for, given that they might predispose people (particularly women) to downplay their personal experiences of discrimination (Kobrynowicz & Branscombe, 1997)." (Cameron, 2001, p. 751)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"**Overview of Regression Analyses** Three sets of hierarchical regression analyses were conducted, with perceived personal discrimination, perceived group discrimination, and the personal/ group discrimination discrepancy as dependent variables. Following procedures described by Aiken and West (1991), gender was dummy-coded (*men* = 0; *women* = 1) and entered along with the centered continuous variables—the three components of social identification, neosexism, sex-role ideology, self-esteem, and social desirability—at Step 1 of the regressions. Two-way interactions involving gender, neosexism, and the social identity variables were tested hierarchically (see Cohen & Cohen, 1983) at Step 2, and the three-way interactions were tested at Step 3. Unstandardized regression coefficients are reported throughout, given their interpretability in the context of interactive effects (Aiken & West, 1991)." (Cameron, 2001, p. 754-755) "Higher levels of personal and group discrimination tended to be perceived by high-neosexism men and low-neosexism women. The centrality of gender identification was positively related to men's personal-level perceptions of discrimination, whereas effects of the emotional facets of social identity—ingroup ties and ingroup affect—occurred jointly with both gender and modern sexism." (Cameron, 2001, p. 743)

Fallacious interpretation of (lack of) statistical significance.	2	"Finally, although the results replicated the personal/group discrimination discrepancy among both women and men, the discrepancy itself was not significantly accounted for by the combined predictor variables." (Cameron, 2001, p. 762) "In summary, with the exception of men's centrality of gender, there is little evidence from the zero-order correlations that social identification is associated with perceptions of discrimination at either the personal or group level. Of primary interest, however, was whether the social identity variables interacted with neosexism and gender; these questions were addressed using the regression analyses reported below." (Cameron, 2001, p. 754) "In summary, analyses of perceived group-level discrimination indicated (a) no evidence of hypothesized main effects of social identity, (b) support for the expected joint effect of gender and neosexism, and (c) mixed support for higher order effects involving gender, neosexism, and social identity." (Cameron, 2001, p. 758)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	2	"**The Personal/Group Discrimination Discrepancy** A 2 (sex) x 2 (type of discrimination: personal/group) ANOVA, with repeated measures on the second factor, was conducted to examine the personal/group discrimination discrepancy. The main effect of sex was significant, indicating that women perceived greater discrimination—averaged across the personal and group levels—than did men, $F(1, 319) = 22.20$, $p < .001$. The within-subjects effect, which replicates the personal/group discrimination discrepancy, was also significant, $F(1, 319) = 47.42$, $p < .001$. These main effects were subsumed within a significant Sex x Type of Discrimination interaction, $F(1, 319) = 5.23$, $p < 05$. Simple effects tests indicated that the personal/group discrepancy was larger for women, $F(1, 319) = 58.67$, $p < .001$, than for men, $F(1, 319) = 8.24$, $p < .01$. The pattern of means in Table I indicates, in concert with the results of the regression analyses reported above, that this is attributable primarily to greater perceptions of group-level discrimination on the part of women. Personal/group discrepancy scores were computed for each participant by subtracting perceived personal discrimination from perceived group-level discrimination, and were regressed on the predictor variables using the same hierarchical procedure as

> for the separate personal and group components. At Step 1, the personal/group discrimination discrepancy was not reliably predicted by the combined independent variables, $F(8, 311) = 1.90$, $p < .06$, $R^2 = .05$, although two effects were significant: sex, $B = .49$, $t(311) = 2.57$, $p = .01$, as reported above, and sex-role ideology, $B = -.34$, $t(311) = -2.00$, $p < .05$, indicating that the discrepancy tended to be smaller for those with relatively liberal sex-role beliefs. None of the interactions entered at Step 2, $F(15, 304) = 1.81$, $p < .05$, $R^2 = .08$, and Step 3, $F(18, 301) = 1.68$, $p < .05$, $R^2 = .09$, was significant." (Cameron, 2001, p. 759)
>
> "Correlations and descriptive statistics involving the social identity measures, the neosexism scale, and perceptions of discrimination are presented in Table I.[2]" (Cameron, 2001, p. 753)
>
> The second footnote: "Although not the focus of the analyses, gender differences on the measures of social identity and gender-related ideology are also of interest, and serve as a reminder of the intergroup background of this investigation. A MANOVA conducted on the social identification subscales yielded a significant multivariate effect of sex, $F(3, 317) = 4.46$, $p < 01$. Univariate tests indicated that the pattern of gender differences on the subscales replicated previous research (Cameron & Lalonde, 2001); that is, although the affective evaluation of group membership was equally positive for members of both sexes, $F(1, 319) = 1.54$, *ns*, women perceived greater ingroup ties than did men, $F(1, 319) = 4.70$, $p < 05$, and indicated that gender was more central to thought and self-definition, $F(1, 319) = 5.51$, $p < 05$; see Table I. A second MANOVA, conducted on the two measures of gender-related beliefs, also yielded a significant multivariate effect of sex, $F(2, 317) = 35.92$, $p < .001$. Not surprisingly, compared to men, women had lower mean levels of neosexism, $F(1, 318) = 59.10$, $p < 001$, and more liberal sex-role beliefs, $F(1, 318) = 64.35$, $p < 001$." (Cameron, 2001, p. 753)

In Figure C9, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(1, 319) = 5.23$; $F(8, 311) = 1.90$; $F(15, 304) = 1.81$; $F(18, 301) = 1.68$.

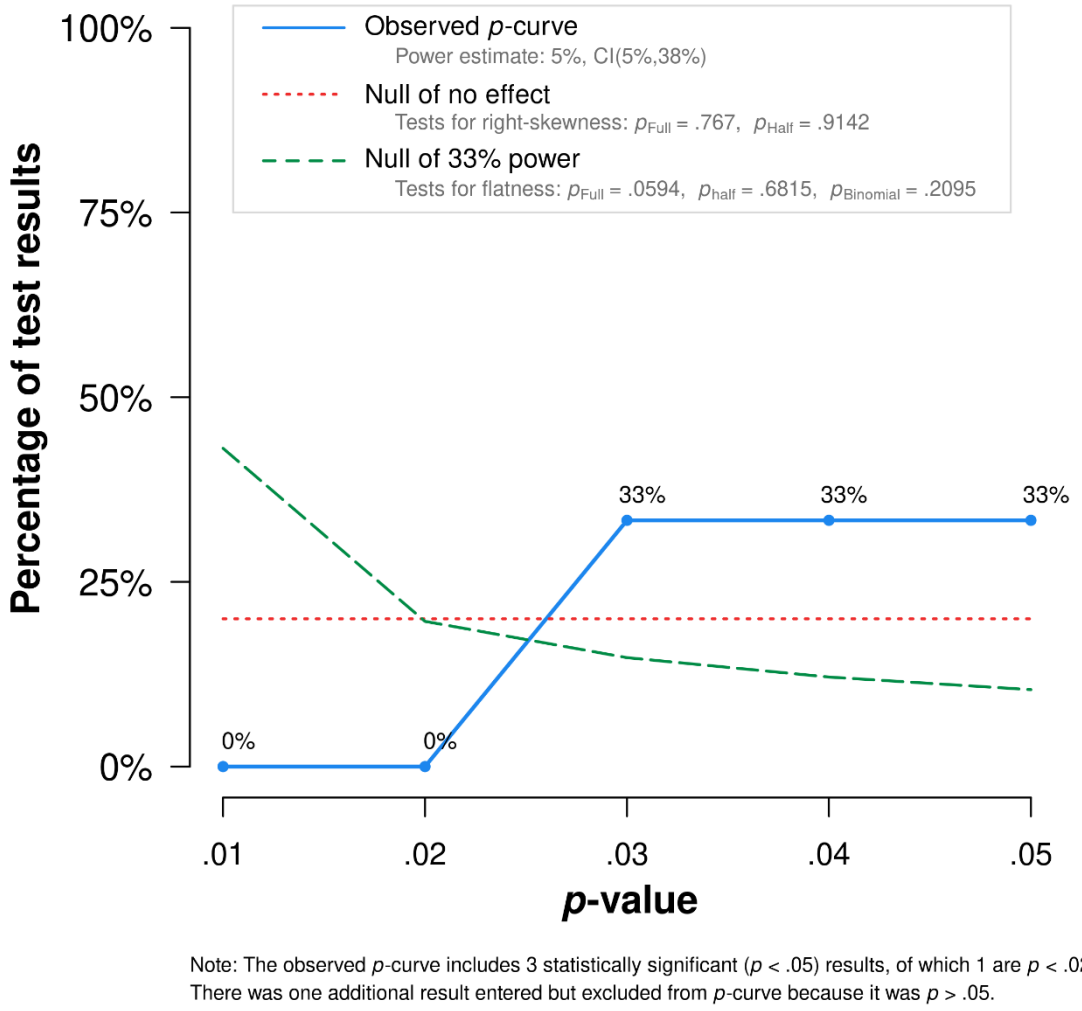


Figure C9. The single-article *p*-curve for the number 7 of the center 10 (i.e., Cameron, 2001).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C10, not only is the half *p*-curve test ($p = .9142$) not significantly right-skewed ($p < .05$), but also both the half ($p = .9142$) and full test ($p = .767$) are not significantly right-skewed ($p < .1$), which implies that the study does not contain evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C10, the 33% power test is $p = .0594$ for the full *p*-curve, for the half p-curve is $p = .6815$, and for the binomial 33%

power test is $p = .2095$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.875	Z=0.73, p=.767	Z=1.37, p=.9142
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p=.2095	Z=-1.56, p=.0594	Z=0.47, p=.6815
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 5% 90% Confidence interval: (5% , 38%)		

Figure C10. Additional statistics for the single-article *p*-curve for the number 7 of the center 10 (i.e., Cameron, 2001).

Number 8 of the Center 10

The number 8 of the center 10 is the first study reported in the paper *Acute Thoughts, Exercise Consistency, and Coping Self-Efficacy* (Gyurcsik et al., 2002). Table C8 shows the score on each of the eight selected RDF for the number 8 paper.

Table C8

Coding paper nr. 8 of the center 10 (i.e., Gyurcsik et al., 2002)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "One purpose of the present study is to determine whether exercisers, classified with respect to tone of acute thinking (i.e., mainly positive or negative) and adherence consistency (i.e., consistent or inconsistent), differ on coping self-efficacy, pre-coping decisional struggle, exercise intention, and post-coping affect. Significant main effects are expected. Consistent exercisers are expected to have significantly (a) higher coping self-efficacy, intention, and affect; and (b) lower decisional struggle than inconsistent exercisers. Positive thinkers are expected to exhibit a similar pattern of scores for these same dependent variables, compared to negative thinkers. Because of the preliminary nature of this study, no interaction effect hypotheses are advanced. A second purpose is to determine whether coping self-efficacy is related to decisional struggle, exercise intention,

		and affect. Based on self-efficacy theory (Bandura, 1997), coping self-efficacy is expected to significantly predict each of these dependent variables." (Gyurcsik et al., 2002, 2138) "One study purpose was to determine whether individuals classified with respect to consistency of exercise adherence and acute thinking tone differed on coping self-efficacy, decisional struggle, exercise intention, and affect. A second study purpose was to examine whether coping self-efficacy predicted struggle, intention, and affect." (Gyurcsik et al., 2002, 2134) "the hypotheses of interest in the current study." (Gyurcsik et al., 2002, 2144)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	3	"A small number of participants (*n* = 22) returned the questionnaire within 2 days of administration in a designated drop box located in the fitness area. There were no anomalous questionnaire responses when this group was compared to the remainder of the sample." (Gyurcsik et al., 2002, 2141)
Sample size (predetermined or not).	2	"Individuals who were active members of community-based fitness clubs or who were actively participating in fitness classes at various universities were approached after a designated exercise session for study participation. Individuals who agreed to participate completed a questionnaire at this time. Each questionnaire was comprised of questions pertaining to demographic information, their prior 4-month pattern of exercise, and the measures as described in the previous section." (Gyurcsik et al., 2002, 2141) "Participants were 160 (140 women, 20 men) healthy people between the ages of 14 and 74 years (*M* = 25.6 years, *SD* = 8.99 years). The sample was composed of students (*n* = 114); people from professional, managerial, technical, and clerical occupations (*n* = 35); and homemakers (*n* = 11). At the time of data collection, all participants were exercising in community-based fitness clubs (n = 92) or in university-based fitness settings (*n* = 68). Their exercise sessions included aerobic exercise (i.e., fitness classes or cardiovascular exercise

		machines) and weight training (n = 106), fitness classes (n = 43), weight training (n = 7), or use of cardiovascular exercise machines (n = 4). The majority of the participants were female, which is representative of typical fitness-club demographics (Dawson, Brawley, & Maddux, 2000).” (Gyurcsik et al., 2002, 2138) “Participants were 160 healthy people (M_{age} = 25.6 years) exercising in fitness settings.” (Gyurcsik et al., 2002, 2134)
Sharing/Openness (i.e., materials, data, code).	2	Measures are fully described on pp. 2139-2140 (Gyurcsik et al., 2002). “Social cognitive, affect, and exercise consistency measures were obtained concurrently.” (Gyurcsik et al., 2002, 2134)
Using covariates and reporting the results with and without the covariates.	3	“The first dummy variable was type of thinker (i.e., positive or negative), and the second dummy variable was adherence consistency (i.e., consistent or inconsistent). These variables were entered before coping self-efficacy to control for any effects that they exerted on the criterion variable. Thus, the initial influence of positive or negative thinking and exercise consistency were controlled in order to examine the added and independent influence of coping self-efficacy on the criterion variable.” (Gyurcsik et al., 2002, 2146)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	1	“In order to investigate if differences existed in the social cognitions and affect of participants classified with respect to the tone of acute thinking and adherence consistency, a 2 x 2 between-subjects MANOVA was conducted. The independent variables were type of thinker (mainly positive or negative) and adherence consistency (consistent or inconsistent) and the dependent variables were pre-coping decisional struggle, coping self-efficacy, post-coping affect, and exercise intention. Assumptions underlying the use of MANOVA were met as indicated by the nonsignificant Levene’s and Box’s tests (ps > .05).” (Gyurcsik et al., 2002, 2143-2144) “A multivariate analysis” (Gyurcsik et al., 2002, 2134) “A one-way between-subjects MANOVA was conducted in which the days per week and weekly 4-month participation rates were compared for the

		inconsistent and consistent exercise groups." (Gyurcsik et al., 2002, 2142) "In order to conduct the 2 x 2 (Type of Thinker x Exercise Consistency) MANOVA and hierarchical multiple regressions that were used to address study purposes, participants were categorized as positive or negative thinkers and consistent or inconsistent exercisers. When categorizing positive or negative thinkers, a two-step procedure was employed. First, a protocol developed by Gyurcsik and Brawley (2000) was used in which participants were classified into one of three groups based on their overall acute thought frequency. (…) Second, positive and negative thinkers were further categorized into two groups-extreme positive thinkers and extreme negative thinkers-because individuals most likely to exhibit characteristic differences in their social cognitions and affect would be those most extreme in their acute thinking. Similar to past research (Gyurcsik & Brawley, 2000), these extreme groups were used in all of the analyses. To obtain extreme groups who had a clear majority of either positive or negative thoughts, cut-off values of -2 (i.e., negative) or +2 (i.e., positive) were chosen. This procedure resulted in 64 clearly negative thinkers and 64 clearly positive thinkers. A *t* test indicated that these two groups differed significantly on overall thought frequency, $t(107) = 18.89$, $p < .0001$." (Gyurcsik et al., 2002, 2141)
Fallacious interpretation of (lack of) statistical significance.	0	η^2 are reported on p. 2144 (Gyurcsik et al., 2002) "A multivariate analysis indicated that positive thinkers experienced significantly lower decisional struggle and higher coping self-efficacy compared to negative thinkers. Further, consistent exercisers experienced significantly lower decisional struggle and higher coping self-efficacy, intention, and positive affect compared to inconsistent exercisers. Regression analyses indicated that coping self-efficacy significantly predicted decisional struggle and intention." (Gyurcsik et al., 2002, 2134) "In sum, these hierarchical multiple regression analyses provided partial support for study hypotheses. The hypotheses that coping self-efficacy would significantly predict decisional struggle and exercise intention were supported. In contrast, the hypothesis that coping self-efficacy

		would significantly predict affect was not supported." (Gyurcsik et al., 2002, 2147) "series of hierarchical multiple regression analyses were conducted to examine the hypotheses of interest in the current study." (Gyurcsik et al., 2002, 2145)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"The overall MANOVA was significant, $F(2, 110) = 57.23$, Pillai's trace = .51, $p < .0001$. Follow-up univariate *F* tests revealed that the inconsistent group exercised on significantly fewer days per week, $F(1, 111) = 27.66$, $p < .0001$; and fewer weeks during the prior 4 months, $F(1, 111) = 108.59$, $p < .0001$, compared to the consistent group. Thus, clear differences in recent past behavior that could determine current beliefs and thoughts about exercise were evident. These two classification procedures resulted in the following groups: (a) positive/consistent, $n = 32$; (b) positive/inconsistent, $n = 32$; (c) negative/consistent, $n = 20$; and (d) negative/inconsistent, $n = 44$. These groupings resulted in unequal numbers in some categories, and this was taken into account in the following 2 x 2 (Type of Thinker x Exercise Consistency) MANOVA analysis. It is important to note that this approach maintained a sufficient number of participants in each group to investigate the research questions (Tabachnik & Fidell, 1996) and permitted comparison with previously published research." (Gyurcsik et al., 2002, 2142) "A significant main effect was found for adherence consistency, $F(4, 121) = 4.34$, Pillai's trace = .13, $p < .003$. Subsequent univariate *F* tests revealed that consistent exercisers had significantly higher coping self-efficacy, $F(1, 124) = 7.05$, $p < .01$ (power = .75, $\eta^2 = .05$); post-coping affect, $F(1, 124) = 4.83$, $p < .03$ (power = .59, $\eta^2 = .04$); exercise intention, $F(1, 124) = 4.70$, $p < .03$ (power = .58, $\eta^2 = .05$); and significantly lower pre-coping decisional struggle, $F(1, 124) = 7.05$, $p < .01$ (power = .75, $\eta^2 = .05$) than did inconsistent exercisers (see Table 2 for estimated marginal means). These results supported study hypotheses. A significant main effect was also found for type of thinker, $F(4, 121) = 5.23$, Pillai's trace = .14, $p < .001$. Subsequent univariate *F* tests revealed that positive thinkers had significantly higher coping self-efficacy, $F(1, 124) = 7.22$, $p < .01$ (power = .76, $\eta^2 = .06$); and lower pre-coping decisional

struggle, $F(1, 124) = 20.16$, $p < .001$ (power = .99, $\eta^2 = .14$), than did negative thinkers (see Table 2 for estimated marginal means). These findings supported study hypotheses. However, in contrast to study hypotheses, positive and negative thinkers did not differ significantly on post-coping affect, $F(1, 124) = 0.08$, $p > .10$ (power = .06, $\eta^2 = .001$) and exercise intention, $F(1, 124) = 2.29$, $p > .10$ (power = .32, $\eta^2 = .02$).

No significant multivariate interaction was found between adherence consistency and type of thinker, $F(4, 121) = 0.83$, Pillai's trace = .03, $p > .10$. Although an interaction between these two variables would seem to follow from theory, the data did not support such a finding. While examination of the raw means of coping self-efficacy, pre-coping struggle, post-coping affect, and intention for the two groups that would be expected to be most different (i.e., positive consistent and negative inconsistent exercisers) were in the expected directions,[4] the differences and possibly the power to detect effects reliably might not have been sufficient." (Gyurcsik et al., 2002, 2144)

"*Pre-coping decisional struggle*. Type of thinker, adherence consistency, and coping self-efficacy were regressed on pre-coping struggle. As seen in Table 3, the overall model was significant, $F(3, 124) = 21.23$, $p < .0001$. As expected, coping self-efficacy was a significant predictor of pre-coping struggle ($\Delta R^2 = .12$) after the significant influence of type of thinker and of adherence consistency were controlled.

Post-coping affect. As seen in Table 3, type of thinker, adherence consistency, and coping self-efficacy were regressed on post-coping affect. The overall model was not significant, $F(3, 124) = 1.64$, $p > .05$. Contrary to expectations, coping self-efficacy was not a significant predictor of affect.

Exercise intention. When type of thinker, adherence consistency, and coping self-efficacy were regressed on exercise intention, the overall model was significant, $F(3, 124) = 9.12$, $p < .0001$ (Table 3). As expected, coping self-efficacy was a significant predictor of intention ($\Delta R^2 = .11$) in addition to the variance accounted for by type of thinker and adherence consistency.

Since pre-coping struggle was significantly associated with exercise intention, this variable was added to the predictive model in order to determine

> whether it explained significant independent variance over and above that accounted for by coping self-efficacy. Although this overall model was significant, $F(4, 123) = 7.69$, $p < .0001$ (adjusted $R^2 = .17$), pre-coping struggle did not contribute significant variation ($\Delta R^2 = .02$, $p > .05$).” (Gyurcsik et al., 2002, 2146-2147)

In Figure C11, the results are shown of entering the following statistics into the online *p*-curve app (“*P*-curve app 4.06,” 2017): $F(2, 110) = 57.23$; $F(4, 121) = 0.83$; $F(3, 124) = 21.23$; $F(3, 124) = 1.64$; $F(3, 124) = 9.12$; $F(4, 123) = 7.69$.

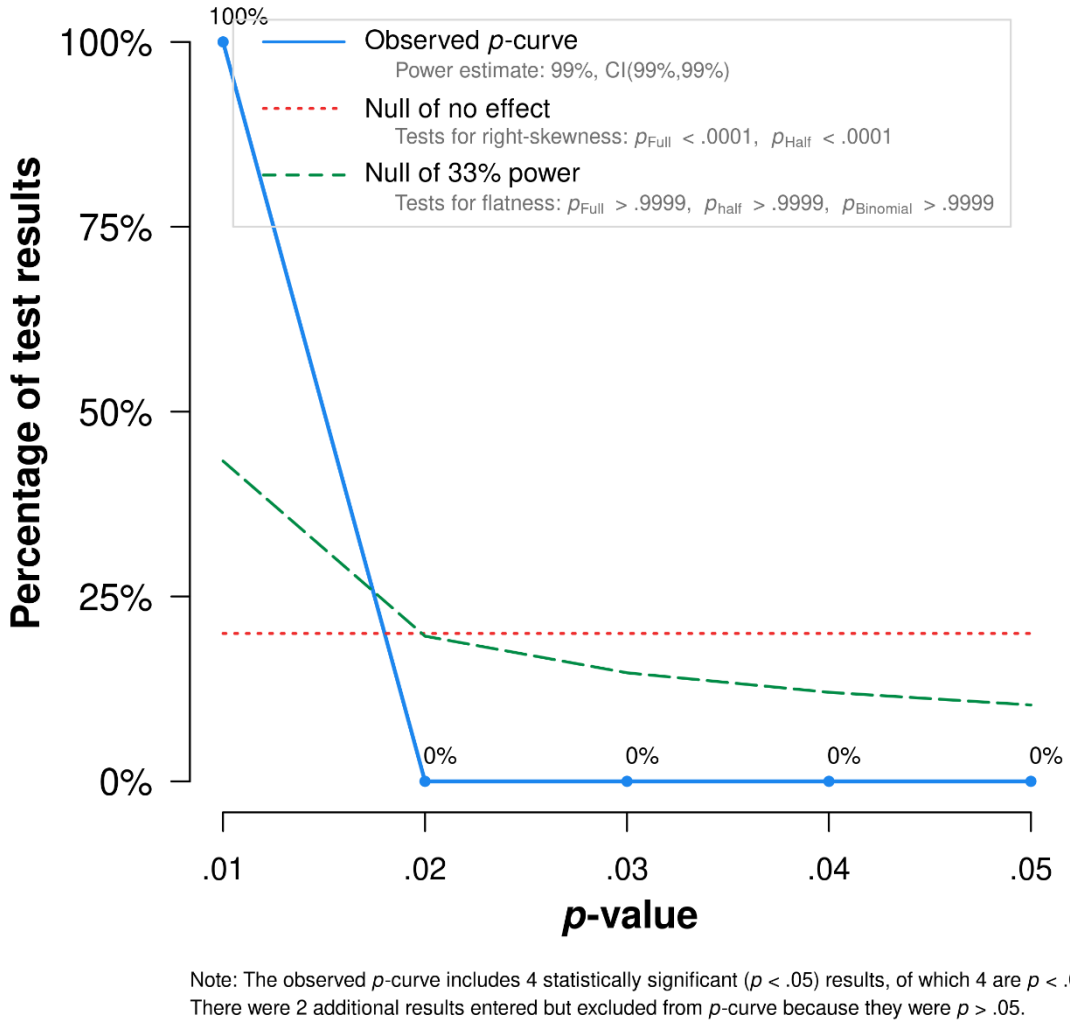


Figure C11. The single-article *p*-curve for the number 8 of the center 10 (i.e., Gyurcsik et al., 2002).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C12, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C12, the 33% power test is $p > .9999$ for the full *p*-curve, for the half p-curve, and for the binomial 33% power test; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.0625	*Z*=-10.32, *p*<.0001	*Z*=-10.03, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*>.9999	*Z*=7.96, *p*>.9999	*Z*=8.35, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure C12. Additional statistics for the single-article *p*-curve for the number 8 of the center 10 (i.e., Gyurcsik et al., 2002).

Number 9 of the Center 10

The number 9 of the center 10 is the first study reported in the paper *Tyramine, a new clue to disinhibition and sensation seeking?* (Thieme & Feij, 1985). Table C9 shows the score on each of the eight selected RDF for the number 9 paper.

Table C9

Coding paper nr. 9 of the center 10 (i.e., Thieme & Feij, 1985)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis	2	"The main hypothesis of this study" (Thieme & Feij, 1985, p. 351) "The aim of this research was to study the relationship between disinhibition (and sensation

planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)		seeking) and urinary MHPG excretion in a larger sample. In addition to MHPG and MHPG-sulphate we measured urinary excretion of free tyramine and the serotonergic metabolite 5-HIAA (5hydroxy-indol-acetic acid) to control for dietary influences." (Thieme & Feij, 1985, p. 349) "We studied the differences in urinary excretion of MHPG, MHPG-sulphate, free tyramine, 5-HIAA and creatinine, between 12 high and 13 low disinhibitors during a low-catecholamine diet." (Thieme & Feij, 1985, p. 349)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"In the MHPG-sulphate group 4 *S*s had to be omitted because of low recovery values." (Thieme & Feij, 1985, p. 350)
Sample size (predetermined or not).	3	"154 introductory psychology and physical education students." (Thieme & Feij, 1985, p. 350) "Twenty-five students were selected from the larger sample on the basis of their Dis scores. Twelve high Dis students (scoring in the upper quartile; 5 males and 7 females) and 13 low Dis students (scoring in the lower quartile; 6 males and 7 females) were selected as Ss in this study. A Gen sensation seeking score was also calculated for the Ss. The median score in the larger sample was used to divide the selected group in highs (n = 14) and lows (n = 11) on general sensation seeking. All Ss volunteered and were paid for their participation." (Thieme & Feij, 1985, p. 350)
Sharing/Openness (i.e., materials, data, code).	2	"A Dutch nonforced-choice Sensation Seeking Scale (SBL; Feij and van Zuilen, 1984) was administered to 154 introductory psychology and physical education students. This questionnaire was constructed after the model of the Zuckerman Sensation Seeking Scales (Forms IV and V; Zuckerman, 1979); it measures the four dimensions of sensation seeking: '*thrill and adventure seeking'* (TAS), '*experience seeking'* (ES), '*boredom susceptibility'* (BS) and '*disinhibition'* (Dis). Satisfactory internal consistencies (cc-coefficients

of 0.80, 0.74, 0.73 and 0.78, respectively) and low to moderate intercorrelations among the scales (ranging from 0.03 and 0.45) have been reported (Feij et al., 1982). In addition to scores on the four subscales, a general (Gen) sensation seeking score is obtained by summing these scores divided by the numbers of items in the respective scales. Several studies on the construct validity of the scales (Feij and van Zuilen, 1984; Feij, Orlebeke, Gazendam and van Zuilen, 1985) yielded results which confirm the research summarized by Zuckerman (1979).

Twenty-five students were selected from the larger sample on the basis of their Dis scores. Twelve high Dis students (scoring in the upper quartile; 5 males and 7 females) and 13 low Dis students (scoring in the lower quartile; 6 males and 7 females) were selected as *S*s in this study. A Gen sensation seeking score was also calculated for the Ss. The median score in the larger sample was used to divide the selected group in highs (n = 14) and lows (n = 11) on general sensation seeking. All *S*s volunteered and were paid for their participation.

About 9 months after the administration of the SBL, *S*s were asked to complete the questionnaire for the second time. The stability of the scores was high: test-retest reliabilities were 0.87, 0.83, 0.79, 0.91 and 0.89 for TAS, ES, BS, Dis and Gen sensation seeking, respectively. Means and standard deviations of the scores were not substantially different on both occasions. The sensation seeking scores referred to in the Results section are the average scores over the two occasions.

In addition to the SBL (and retest) *S*s completed a Dutch version of the Zung Depression Scale (Zung, 1965) and provided information about their smoking habits (i.e. the estimated number of cigarettes, cigars and/or pipes per day).

Procedure

After completion of the questionnaires. each *S* received a written instruction for the method of urine collection. *S*s were asked to collect 24 hr urine while using a low-catecholamine diet, and not to overinvolve in sports, stressful situations, smoking and alcohol or coffee consumption.

None of the *S*s received medication at the time of this study. Nine of the 14 female *S*s used oral contraceptives. We did not account for this as an independent variable. Urine was collected in plastic containers with 3 ml hydrochloric acid (6 N HCl)

		and *S*s were asked to keep the container in a refrigerator or at a cool spot, and to return the sample as soon as possible. Then the containers were stored at -20°C till determination within a few weeks. *Urine analysis* Total MHPG excretion was measured as free MHPG with a gas-liquid chromatographic (GLC) method after enzymatic deconjugation. MHPG-sulphate was also measured as free MHPG with this GLC method after separation and deconjugation with sulphatase according to Eichholtz, Binkhuyzen and Thieme (1984). Results were expressed as mg/g creatininej24 hr. Tyramine was estimated with a fluorimetric method according to Udenfriend (1962) and expressed as ,LLg/g creatinine/24 hr. 5-HIAA was estimated with a fluorimetric method according to Korf and Valkenburgh-Sikkema (1969). and expressed as mg/g creatinine/24 hr." (Thieme & Feij, 1985, p. 350) "Disinhibition (a subdimension of sensation seeking) was measured by a reliable and valid Dutch Sensation Seeking Scale." (Thieme & Feij, 1985, p. 349)
Using covariates and reporting the results with and without the covariates.	0	"In addition to MHPG and MHPG-sulphate we measured urinary excretion of free tyramine and the serotonergic metabolite 5-HIAA (5hydroxy-indol-acetic acid) to control for dietary influences." (Thieme & Feij, 1985, p. 349) "We were surprised to find a significant difference in free tyramine excretion. High sensation seekers and disinhibitors excrete more tyramine, and also when no correction for creatinine excretion (body weight) is made this difference is still present: t = 2.21 and 2.81. respectively ($P < 0.01$). Without creatinine correction, tyramine, MHPG and MHPG-sulphate were significantly interrelated. This is what Linnoila, Karoum and Potter (1982) described in depressive patients. When corrected for creatinine excretion these interrelations disappeared in our study." (Thieme & Feij, 1985, p. 352) "Nor did tyramine, uncorrected for creatinine excretion [which of course was higher for male than for female *S*s; $t(23) = 3.83$, $P < 0.001$], differentiate between the sexes." (Thieme & Feij, 1985, p. 352)

Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	
Fallacious interpretation of (lack of) statistical significance.	2	"The results showed a significant difference between both groups for tyramine only. High disinhibitors (sensation seekers) excreted more tyramine than the lows. Possible explanations for this unpredicted finding are discussed." (Thieme & Feij, 1985, p. 349)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"We were surprised to find a significant difference in free tyramine excretion. High sensation seekers and disinhibitors excrete more tyramine, and also when no correction for creatinine excretion (body weight) is made this difference is still present: t = 2.21 and 2.81. respectively ($P < 0.01$). Without creatinine correction, tyramine, MHPG and MHPG-sulphate were significantly interrelated. This is what Linnoila, Karoum and Potter (1982) described in depressive patients. When corrected for creatinine excretion these interrelations disappeared in our study." (Thieme & Feij, 1985, p. 352) "As a matter of fact, the difference in mean tyramine output between male *S*s ($\bar{X}$ = 310.00. SD = 83.34) and female *S*s ($\bar{X}$ = 401.43, SD = 186.70) was not significant: $t(18.85) = 1.64$ (NS)." (Thieme & Feij, 1985, p. 352) "Nor did tyramine, uncorrected for creatinine excretion [which of course was higher for male than for female *S*s; $t(23) = 3.83$, $P < 0.001$], differentiate between the sexes." (Thieme & Feij, 1985, p. 352)

In Figure C13, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $t(18.85) = 1.64$; $t(23) = 3.83$.

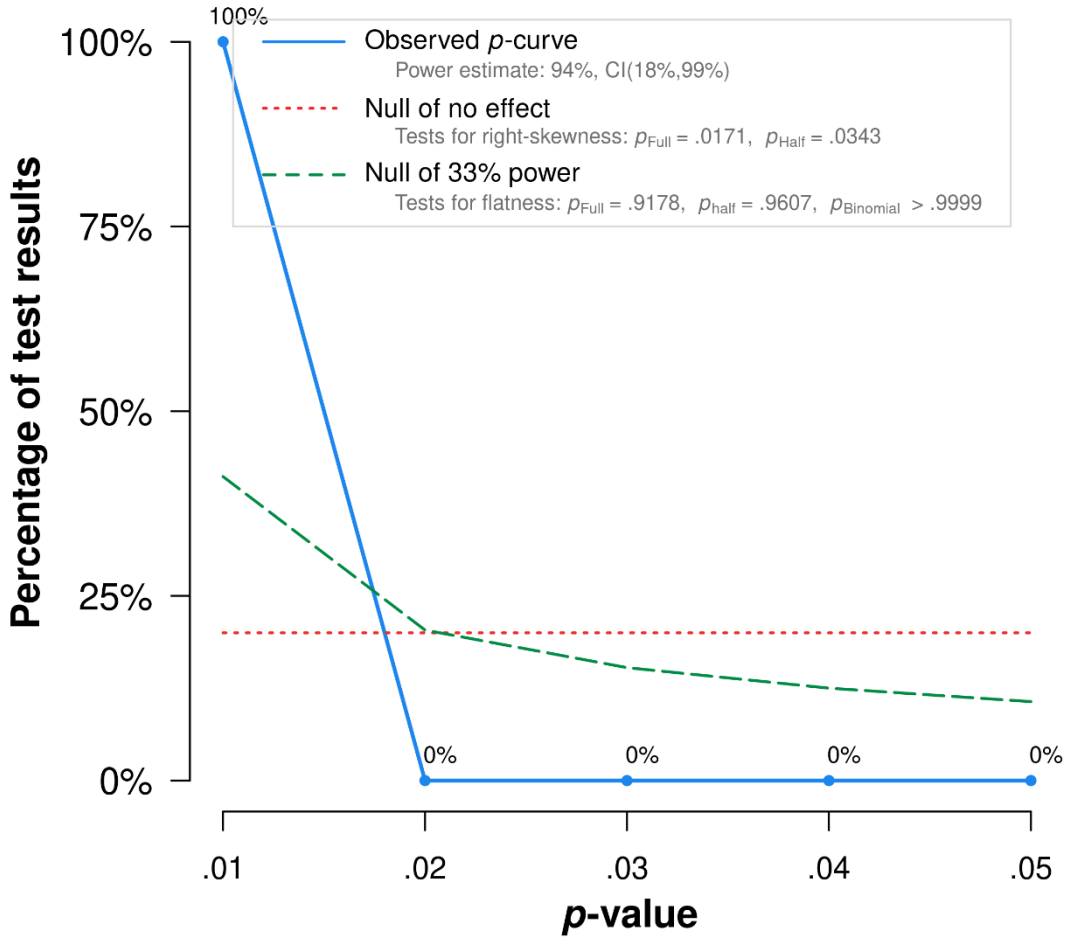


Figure C13. The single-article *p*-curve for the number 9 of the center 10 (i.e., Thieme & Feij, 1985).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C14, not only is the half *p*-curve test ($p = .0343$) significantly right-skewed ($p < .05$), but also both the half ($p = .0343$) and full test ($p = .0171$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C14, the 33% power test is $p = .9178$ for the full *p*-curve, for the half p-curve is $p = .9607$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	Continuous Test *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. *(Right skew)*	p=.5	Z=-2.12, p=.0171	Z=-1.82, p=.0343
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p>.9999	Z=1.39, p=.9178	Z=1.76, p=.9607
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 94% 90% Confidence interval: (18% , 99%)		

Figure C14. Additional statistics for the single-article *p*-curve for the number 9 of the center 10 (i.e., Thieme & Feij, 1985).

Number 10 of the Center 10

The number 10 of the center 10 is the first study reported in the paper *The Effects of Race, Weight, and Gender on Evaluations of Writing Competence* (Surmann, 1997). Table C10 shows the score on each of the eight selected RDF for the number 10 paper.

Table C10

Coding paper nr. 10 of the center 10 (i.e., Surmann, 1997)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	3	My purpose in the present study was to expand the traditional American gender bias research paradigm to include the commonly occurring realistic human covariates of weight and race to investigate the interactive prejudicial potential of target race, weight, and gender on participants' ratings of writing competence. Although studies conducted in the United States and Australia confirmed that overweight people are perceived negatively and discriminated against by others (Brink, 1988; DeJong, 1980; Harris & Hopwood, 1982; Jasper & Klassen, 1990a, b; Larkin & Pines, 1979; Rothblum, Brand, Miller, & Oetjen, 1990; Young & Powell, 1985), and previous research manipulated race and gender (Hamner, Kim, Baird, & Bigoness, 1974; Howell, 1983), prior studies neglected to measure the compounding prejudicial effects of target weight, race, and gender on readers' evaluations of writing competence. The present study introduced these characteristic factors (African American/Caucasian, overweight/nonoverweight, male/female) discretely with minimal and essentially identical case information so that relationships among and between these variables could be studied simultaneously. The terms *overweight* and *obesity* are regarded as concepts on a continuum, with obesity defined as 30% over an established preferred weight (Hiller, 1981). The stimuli in the present study depicted targets who were noticeably overweight but not severely obese. The negative conclusions of the obesity research, considered together with Howell's (1983) findings that African American women suffered discrimination based on both race and gender, would predict that overweight African American female authors in the present study could be expected to suffer additive prejudice, producing the lowest evaluations of all target groups. Although the commonly understood concept of race reflects an unscientific method of human differentiation, in the present study I used the concept in the traditional sense for the purposes of analysis. (Surmann, 1997, p. 174)

		ABSTRACT. This study examined the potential interactive effects of target race, weight, and gender on American Caucasian students' ratings of writing competency. Stimulus (Surmann, 1997, p. 173)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	*Participants* The participants were 64 Caucasian/Euro-American students enrolled in introductory psychology at a midwestern U.S university during 1992. The sample included 27 male participants and 37 female participants, all of whom listed their race as "white" or "Caucasian" in response to an open-ended demographic question that requested their race. (Surmann, 1997, p. 175)
Sharing/Openness (i.e., materials, data, code).	3	*Materials* The articles and photographs were combined in a newspaper article format. The articles were excerpts taken from Coon's *Introduction to Psychology* (1986); they gave a general overview of eight human interest topics from various branches of psychology: physiological, educational, social, and developmental. The photographs, taken from various magazines, were selected on the basis of either average attractiveness or quality of realistic representation of each variable group for the age range of 30-49. (Surmann, 1997, p. 175) and gender on American Caucasian students' ratings of writing competency. Stimulus materials were eight articles in a newspaper byline format with photographs of the authors that conveyed the characteristics of race (African American/Caucasian), weight (overweight/nonoverweight), and gender (male/female). The participants were given eight articles and were asked to rate the writing competency of the author on the four qualities of style, clarity, logic, and overall writing ability/competency. The results showed that the (Surmann, 1997, p. 173)

Procedure

Participants read an instruction sheet stating that the purpose of the experiment was "to explore the analytical thought processes that people use to evaluate written passages of text." The stimulus materials were presented as student articles with photo-bylines taken from the psychology department's section of a fictitious college newspaper. Except for the author's name and photograph, no personalized information about each target was provided. A portion of the instruction sheet was designed to function as a distracter, informing the participants of the study's pseudo-purpose—to rate the authors' writing competency—and providing additional information about various aspects of general reading comprehension. Participants were asked to rate the attached articles on the four characteristics: style, clarity, logic, and overall writing ability/competency; however, these characteristics were not explicitly defined in the instructions.

To convey the variables of weight. race, and gender, the eight photographs included a nonobese Caucasian woman, a nonobese Caucasian man, a nonobese African American woman, a nonobese African American man, an obese Caucasian woman, an obese Caucasian man, an obese African American woman, and an obese African American man. Counterbalancing required that every photograph be tested with every article, which resulted in 64 different combinations of author and article. The participants were given eight articles in a random order based on a random number table. They were instructed to rate the writing competency of the author for style, clarity, logic, and overall writing ability/competency on a 5-point scale ranging from *unfavorable* (1) to *favorable* (5). Information about participants' age, gender, and race was requested on the last page of the survey. The study was then a 2(participant gender) × 2(author gender) × 2(author race) × 2(author weight) design.

(Surmann, 1997, p. 175-176)

Using covariates and reporting the results with and without the covariates.

3

My purpose in the present study was to expand the traditional American gender bias research paradigm to include the commonly occurring realistic human covariates of weight and race to investigate the interactive prejudicial potential of target race, weight, and gender on participants' ratings of writing competence. Although studies conducted in the United States and Australia confirmed that overweight people are perceived negatively and discriminated against by others (Brink, 1988; DeJong, 1980; Harris & Hopwood, 1982; Jasper & Klassen, 1990a, b; Larkin & Pines, 1979; Rothblum, Brand, Miller, & Oetjen, 1990; Young & Powell, 1985), and previous research manipulated race and gender (Hamner, Kim, Baird, & Bigoness, 1974; Howell, 1983), prior studies neglected to measure the compounding prejudicial effects of target weight, race and gender on readers' evaluations of writing competence. The present study introduced these characteristic factors (African American/Caucasian, overweight/nonoverweight, male/female) discretely with minimal and essentially identical case information so that relationships among and between these variables could be studied simultaneously. The terms *overweight* and *obesity* are

(Surmann, 1997, p. 174)

Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.

3

tency on a 5-point scale ranging from *unfavorable* (1) to *favorable* (5). Information about participants' age, gender, and race was requested on the last page of the survey. The study was then a 2(participant gender) × 2(author gender) × 2(author race) × 2(author weight) design.

(Surmann, 1997, p. 176)

To control for inherent differences in the quality of writing of the different articles, I adjusted individual article scores with respect to the difference between the mean for each of the eight articles and the overall mean for all articles. Article ratings were then analyzed with a multiple analysis of variance, with repeated measures on author weight, author race, and author gender. An alpha level of .05 was used for all statistical tests.

(Surmann, 1997, p. 176)

Fallacious interpretation of (lack of) statistical significance.	3	showed no prejudicial effect as a function of target gender. The present study supported a body of evidence that American men devalue women (Fidell, 1970; Ward, 1981); however, the men in this study degraded only Caucasian female authors. As the results from Swim et al. (1989) would have predicted, the mean difference was relatively small. To interpret this small, but strongly significant, difference as "negligible," however, as these authors concluded, assumes that the expression of prejudice in the laboratory will be obvious rather than subtle. Furthermore, in a reliable, albeit small, amount of quantitative gender discrimination, the importance and subsequent potential for damage cannot be underestimated. Although the rationale behind the behavior of the male participants remains unknown, the specificity of the discrimination observed in the present experiment could be related to the male participants' beliefs about the status of the Caucasian female authors and the other target groups studied. (Surmann, 1997, p. 177) style, clarity, logic, and overall writing ability/competency. The results showed that the women gave higher style, logic, and overall writing ability/competency ratings than did the men, and overweight authors received higher ratings of logic than nonoverweight authors. The men gave female Caucasian authors lower ratings of clarity than did female participants. (Surmann, 1997, p. 173) **Discussion** Unlike Howell's (1983) finding, African American women were not differentially evaluated on the basis of sex and race, but instead received ratings from male participants that were relatively equal to male authors of both races. (Surmann, 1997, p. 176)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	2	*Results* To control for inherent differences in the quality of writing of the different articles, I adjusted individual article scores with respect to the difference between the mean for each of the eight articles and the overall mean for all articles. Article ratings were then analyzed with a multiple analysis of variance, with repeated measures on author weight, author race, and author gender. An alpha level of .05 was used for all statistical tests. For clarity ratings, the data indicated an interaction effect in which male participants gave female Caucasian authors statistically significant lower ratings than male authors, $F(1, 62) = 6.60$, $p = .013$; in contrast, female participants gave consistent ratings regardless of an author's gender or race. Male participants did not discriminate against male authors of either race or against female African American authors. For African American male authors, the mean clarity ratings were 3.73 by male participants and 3.97 by female participants. For African American female authors, the mean ratings were 3.84 by male participants and 3.93 by female participants. For Caucasian male authors, the mean ratings were 3.79 by male participants and 3.81 by female participants. The mean ratings for Caucasian female authors were 3.45 by male participants and 4.04 by female participants. Overweight authors received statistically significant higher logic ratings than nonoverweight authors, $F(1, 62) = 5.26$, $p = .025$. The mean logic rating was 3.90 for overweight authors, and 3.71 for nonoverweight authors. The mean ratings for style, logic, and overall writing ability/competency indicated that, in comparison with male participants, female participants gave statistically significant higher ratings of style, $F(1, 62) = 8.19$, $p = .006$; logic, $F(1, 62) = 4.78$, $p = .033$; and overall writing ability/competency, $F(1, 62) = 8.91$, $p = .004$. The mean style, logic, and overall writing ability/competency ratings were 3.62, 3.63, and 3.65, respectively, for male participants, and 3.95, 3.93, and 3.98, respectively, for female participants. (Surmann, 1997, p. 176)

In Figure C15, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(1, 62) = 6.60$; $F(1, 62) = 5.26$; $F(1, 62) = 8.19$; $F(1, 62) = 4.78$; $F(1, 62) = 8.91$.

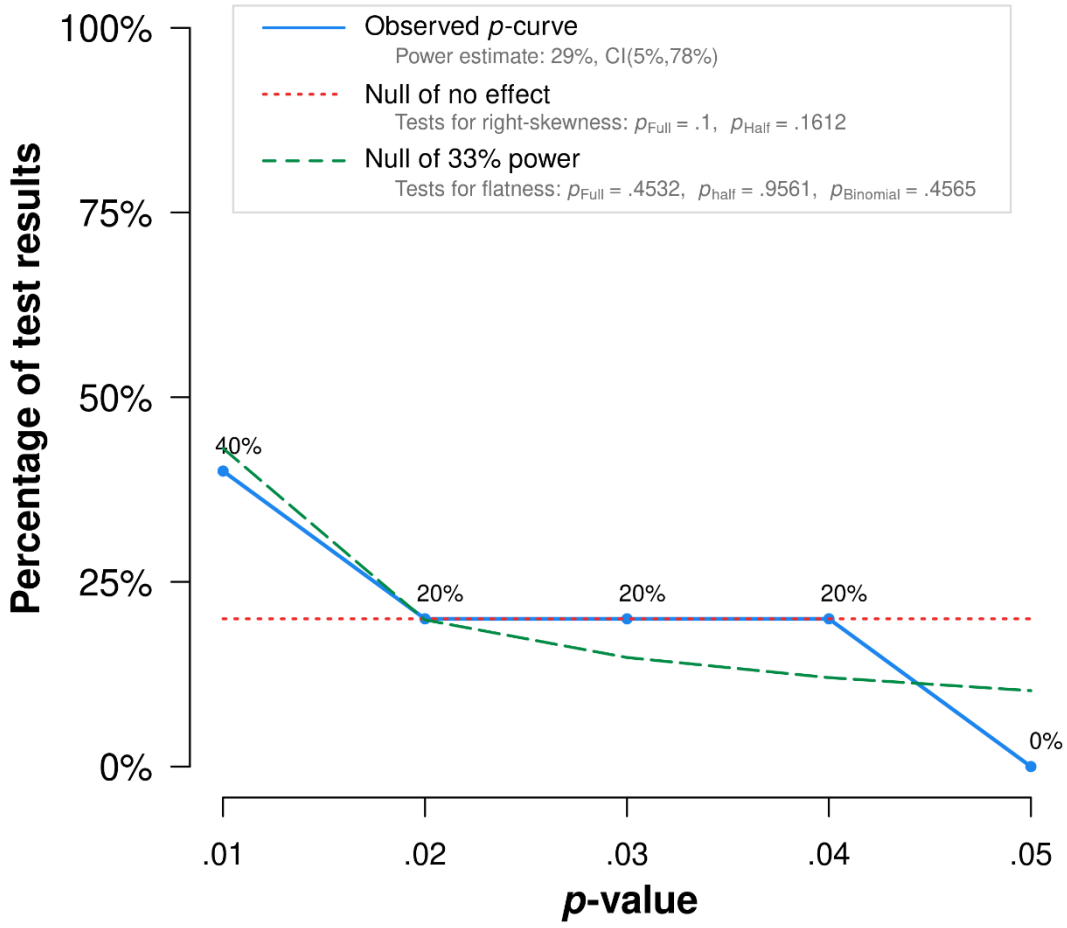


Figure C15. The single-article *p*-curve for the number 10 of the center 10 (i.e., Surmann, 1997).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure C16, not only is the half *p*-curve test ($p = .1612$) not significantly right-skewed ($p < .05$), but also both the half ($p = .1612$) and full test ($p = .1$) are not significantly right-skewed ($p < .1$), which implies that the study does not contain evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full p-curve or both the half p-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure C16, the 33% power test is $p = .4532$ for the full *p*-curve, for the half p-curve is $p = .9561$, and for the binomial 33%

power test is $p = .4565$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	p=.5	Z=-1.28, p=.1	Z=-0.99, p=.1612
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p=.4565	Z=-0.12, p=.4532	Z=1.71, p=.9561
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 29% 90% Confidence interval: (5% , 78%)		

Figure C16. Additional statistics for the single-article *p*-curve for the number 10 of the center 10 (i.e., Surmann, 1997).

Appendix D. Scoring the Bottom 10 Studies on the RDF Checklist

In order to map the DFS of the bottom 10 studies, each of the ten studies is scored on RDF. This Appendix contains the scores for each study on each of the eight items on the RDF checklist.

Number 1 of the Bottom 10

The number 1 of the bottom 10 is the first study reported in the paper *The Persuasive Effects of a Real and Complex Communication* (Puddifoot, 1996). Table D1 shows the score on each of the eight selected RDF for the number 1 paper.

Table D1

Coding paper nr. 1 of the bottom 10 (i.e., Puddifoot, 1996)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	The case study investigated in the present research was an instance in which something akin to a loosely controlled large-scale experiment was conducted on behalf of a government agency. The methodology permitted a measure of control through careful sampling and the standardization of the data collection procedure. The kinds of data that were being collected allowed the effects of an important communication to be monitored unusually closely, although this was by no means the main purpose of the present research. (Puddifoot, 1996, p. 448) The LGC planned to ask representative samples of residents in each area across the nation to indicate their preferred choice from a number of options regarding the reform of local government. After residents had indicated their preferred options, they would be briefed about the details and the cost of each option and subsequently asked to indicate their preferred options again. It was expected that preferences might change in view of the new information. (Puddifoot, 1996, p. 449) Exploratory: In the present study, I used a secondary analysis of two large surveys that were conducted for the LGC, to determine the underlying factors that distinguished the changers (those who changed their views) from the nonchangers. A direct comparison was made with the larger sample from which the changers were in effect self-selected, and which in this respect served as a field control. This procedure contrasts with the general procedure of Hovland et al. (1953), which involved comparisons between participants who were randomly assigned to an experimental or a control condition. (Puddifoot, 1996, p. 449)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to	0	No exclusions.

deal with outliers in an ad hoc manner.		
Sample size (predetermined or not).	1	*Field Research for Durham and Cleveland Counties* The sampling procedure for the LGC public opinion surveys was identical for the two county areas. In both areas, quota samples were secured so as to represent the demographic characteristics of each area, as determined by the 1991 national census. Following are the results for Durham County, with the equivalent figures for Cleveland County immediately afterward in parentheses. In Durham, a sample of 2,478 (1,235) was drawn from a total of 218 (105) sampling points; 49% (48%) were men and 51% (52%) were women. Age quotas were as follows: 18–24 years = 11% (13%); 25–34 years = 20% (21%); 35–44 years = 18% (20%); 45–54 years = 16% (14%); 55–64 years = 4% (14%); 65–74 years = 13% (13%); and 75+ years = 7% (6%). I used the employment status of the head of (Puddifoot, 1996, p. 449)
Sharing/Openness (i.e., materials, data, code).	2	(unskilled) = 20% (20%); E (other/unemployed) = 24% (24%). After establishing contact with a respondent who fit the quota requirements, the interviewers met with the respondent at his or her home, as described previously, and discussed the information that is provided in Appendixes A and B. The two surveys were conducted concurrently during July and August of 1993. (Puddifoot, 1996, p. 450) A relatively large number (59) of the 228 Durham changers altered their original stated preference in favor of the existent system of local government to another preference, but none of the new preferences was clearly favored. Almost half these changers shifted to "don't know" as a result of the new information that was presented in the briefing. More than half (63%) of the respondents who originally preferred Option B (eight authorities) and about 40% of the respondents who originally preferred the other remaining options changed their preferences to the existent system of local government. Thirty-two respondents changed their preferences to Option D (one authority), 27 respondents changed their preferences to Option C (two authorities), 22 respondents changed their preferences to Option E (four authorities), and 11 respondents changed their preferences to Option B (eight authorities). (Puddifoot, 1996, p. 450) One hundred ninety of the 1,235 respondents in the Cleveland sample changed their original preferences for local government reorganization after the briefing. The existent two-tier system (Option A) lost the support of 18 respondents but gained the support of 69 respondents, a net gain of 51. The first of the four authority-solutions (Option B, net gain = 5) and the one-authority solution (Option F, net gain = 10) also demonstrated small gains. Net losses were evident for the second of the four authority-solutions (Option C, net loss = 7), the two-authorities solution (Option D, net loss = 11), and the three-authorities solution (Option E, net loss = 7). Seven residents changed their preferences from the one-authority unitary county solution (Option F) to the existent two-tier system (Option A), despite the fact that the briefing portrayed the one-authority solution as the best option financially. Given the financial disadvantages of Option A relative to the other options, in conjunction with the stated equivalence of the level of services, the shift in the preferences of 69 residents (51 net) to the existent two-tier system was unexpected. (Puddifoot, 1996, p. 451)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical	3	

assumptions in an *ad hoc* manner.		
Fallacious interpretation of (lack of) statistical significance.	2	The LGC reported almost no change in the preferred options of the residents of Cleveland and Durham counties; the highest proportion of net change after the briefing in any district in the two counties was reported to be 1%. However, a case-by-case examination in a secondary analysis indicated that the figures for net overall change masked the extent of individual change. The results of this more detailed analysis, reported in this article, indicated that the 228 residents of Durham County who changed their original preferences actually constituted 9.2% of the overall sample of 2,478. One hundred ninety of the 1,235 respondents in Cleveland county, or 15.4 % of the overall sample, changed their preferences after briefing. (Puddifoot, 1996, p. 450) The net effect of the briefing on the Durham sample was to increase professed support for the existent two-tier system and to reduce support for the eight-authority option. The eight-authority option proposed a separate authority for each of the following areas in the district—Chester-le Street, Darlington, Derwentside, Durham City, Easington, Sedgefield, Teesdale, and Wear Valley—and, according to the briefing, was the option that would involve the least expensive start-up costs. The net effect of the briefing on the other options was almost neutral, with only small changes occurring. (Puddifoot, 1996, p. 451)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	The paper does not disclose enough statistics to calculate the single-article *p*-curve.

Number 2 of the Bottom 10

The number 2 of the bottom 10 is the first study reported in the paper *Does the Medium still Matter? The Influence of Gender and Political Connectedness on Contacting U.S. Public Officials Online and Offline* (Brundidge et al., 2013). Table D2 shows the score on each of the eight selected RDF for the number 2 paper.

Table D2

Coding paper nr. 2 of the bottom 10 (i.e., Brundidge et al., 2013)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide	2	Confirmatory: "*H_1: Both a) offline and b) online, gender (female) is positively related to signing petitions*." (Brundidge et al., 2013, p. 7) "*H_2: Gender (female) is inversely related to contacting public officials directly, via a) sending a letter to public officials and b) emailing public officials*." (Brundidge et al., 2013, p. 7)

between planned and unplanned] or ad hoc)		"*H_3: Political connectedness developed via SNSs is positively related to contacting public officials offline and online, via (a) sending a postal letter to public officials (b) signing pen and paper petitions, (c) emailing public officials, and (d) signing online petitions.*" (Brundidge et al., 2013, p. 8) "*H_4: Political connectedness developed via SNSs is more strongly related to a) emailing public officials than sending a postal letter to public officials and b) signing online petitions than signing paper petitions*." (Brundidge et al., 2013, p. 8) "*H_5: Gender moderates the relationship between SNS connectedness and signing online petitions*."(Brundidge et al., 2013, p. 8)
		"This study employs a secondary analysis of U.S. nationally representative data from the Pew Internet 2008 civic engagement survey (*N*=2251) to examine the degree to which contacting public officials both online and offline is explained by the variables of gender and political connectedness." (Brundidge et al., 2013, p. 3)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	1	"The Pew survey included a filter question, such that only Internet users (*N* = 1,655) answered questions related to online behavior. Those missing values on Internet related items associated with this filter, were therefore recoded as "0" or "never" since we knew that as non-Internet users they could not engage in online activities (i.e., SNS political connectedness, online contact of public officials). Missing values for income were replaced with the mean. All other missing values on all variables were deleted listwise." (Brundidge et al., 2013, p. 10)
Sample size (predetermined or not).	3	"This study is based on a secondary analysis of a Pew Internet dataset created from a U.S. nationally representative survey of adults over the age of 18. The dataset was originally used for the Pew Internet and Civic Engagement study. Princeton Survey Research Associates International conducted the survey between August 12 and August 31, 2008. It employed a random digit dial (RDD) sample of telephone numbers selected from exchanges in the continental United States. The researchers contacted 9,434 people and completed 2,251 surveys (*N* = 2251), a 23.9 % acceptance rate. The data were weighted for response bias across gender, age, and education. The demographic

		weighting parameters were based on a Pew analysis of the most recently available Census Bureau's March 2007 Annual Social and Economic Supplement. The analysis yielded population parameters for the demographic characteristics of adults aged 18 or older, living in continental US households. These parameters were compared with the sample characteristics to construct sample weights (Smith et al. 2009)." (Brundidge et al., 2013, p. 8-9)
Sharing/Openness (i.e., materials, data, code).	2	Materials: "The *age* of respondents was assessed with an item that asked respondents to place themselves in one of six age categories (median=45–54). *Education* was measured by asking respondents to place themselves in one of four education categories (less than high school, high school graduate, some college, and college graduate (median=some college). Income was evaluated by asking respondents to report their total household income for the previous year (2007) by selecting from nine categories ranging from less than $10,000 to $150, 000 or more (median=$40,000 to under $50,000). However, for ease of analysis, we treated age, education, and income into continuous variables. Age was therefore a six-point index (*M*=3.6, *SD*=1.6), education was a four-point index (*M*=2.7, *SD*=1), and income was nine-point index (*M*=5.1, *SD*=2.2). Race was dummy coded with non-White equal to 0 and White equal to 1 (78.8 %) (*M*=.79, *SD*=.41). (…) *Offline political connectedness* was assessed with an index that included five yes/no items (which we dummy coded, 0 no, 1 yes) that asked if respondents had attended a political rally (*M*=.12, *SD*=.33), a political meeting on local, town or school affairs meeting (*M*=.24, *SD*=.43), worked or volunteered for a political e public policy or public, not including a political party (*M*=.15, *SD*=.36), or worked with fellow citizens to solve a problem in their community (*M*=.28, *SD*=.45). (…) *SNS political connectedness* was assessed through the use of a three-item index that asked respondents whether or not they had ever participated in political activities via social networking sites. The index included the following items: whether or not they had started or joined a political group or group supporting a cause on a social networking site (*M*=.03, *SD*=.18), signed up as a "friend" of any candidates on a social networking site (*M*=.03,

		SD=.17), or posted political news for friends or others to read on a social networking site (*M*=.03, *SD*=.16). Each of the three items was measured on a yes or no binary scale (…) *Contacting Public Officials Online and Offline (Criterion variables)* Following Bimber's (1999) general approach, we assessed whether or not respondents contacted public officials, both online and offline. The offline related items asked whether or not respondents had ever (yes or no, dummy coded): contacted a national, state or local public official in person, by phone or by letter about an issue that is important to you; (*M*=.24, *SD*=.43) or signed a paper petition (*M*=.25, *SD*=.43). The online related items asked respondents if they had ever (also yes/no, dummy coded): sent an email to a national, state or local public official about an issue that is important to you (*M*=.18, *SD*=.38) or signed a petition online (*M*=.14, *SD*=.35)." (Brundidge et al., 2013, p. 9-10) "We therefore employ U.S. nationally representative data from the Pew Internet 2008 Civic Engagement Survey to evaluate the persistence of online gender gaps in the realm of one particular form of political participation, citizens' contacting of public officials." (Brundidge et al., 2013, p. 3)
Using covariates and reporting the results with and without the covariates.	3	"We additionally find that that gender moderates the relationship between political connectedness developed via social networking sites and contacting public officials, such that women gain even further advantage in signing online petitions, but also gain further disadvantage in writing a letter/calling public officials and signing offline petitions." (Brundidge et al., 2013, p. 3) "*Gender (Predictor Variable)* For most of our hypotheses, gender is the key predictor variable. For gender (48.8 % female, 51.2 % male), females were dummy coded 1 and males coded 0 (*M*=.51, *SD*=.50)." (Brundidge et al., 2013, p. 9) "*Socio-Demographic Variables (Controls)* Due to their centrality in contacting public officials, and in political participation more generally, sociodemographic variables related to SES, were included in analyses as controls. It is essential to

		disaggregate SES, however. Early studies that examined SES produced contradictory findings (e.g., Coulter 1992; Verba and Nie 1972). Age, for example, is not typically included in a measure of SES, yet is a central to contacting public officials—research has shown that communicating with government is largely a function of education and of age (Rosenstone and Hansen 1993; Verba et al. 1995). We therefore control for demographic variables individually, rather than aggregating them into an overall measure of SES." (Brundidge et al., 2013, p. 9)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"our hypotheses, which we test with a combination of chi-square analysis and logistic regression. The chi-square analyses are used to examine the relationships between our key predictive variables (i.e., gender and online/offline political connectedness) and our criterion variables (i.e., signing petitions, online and offline, emailing public officials, and writing a postal letter to a public official) irrespective of controls. We then use logistic regressions to assess the impact of our predictor variables and interaction terms on our criterion variables while keeping demographic variables (age, income, race, and education) related to socioeconomic status (SES) constant." (Brundidge et al., 2013, p. 4)
Fallacious interpretation of (lack of) statistical significance.	0	Table 2 contains partial η^2 (Brundidge et al., 2013, p. 11). "There were some notable socio-demographic differences among men and women and some relatively minor discrepancies between the Pew sample and the U.S. population. Both chi-square analyses (see Table 1) and an ANOVA (see Table 2) reveal small but significant gender differences in both age and income, with women being slightly younger and earning somewhat less income than men. The income inequality as expressed in mean levels (see Table 2) in the Pew sample is generally consistent with Census estimates, though the Pew sample is generally wealthier than Census median estimates." (Brundidge et al., 2013, p. 9) "Chi Square also gives us a preliminary answer to our first research question. While there are gender gaps in the chi-square analyses favoring men in both online and offline forms of contact, they are importantly slightly smaller for online forms of contact" (Brundidge et al., 2013, p. 11)

Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0

Table 3 Chi-square of gender and contacting government officials

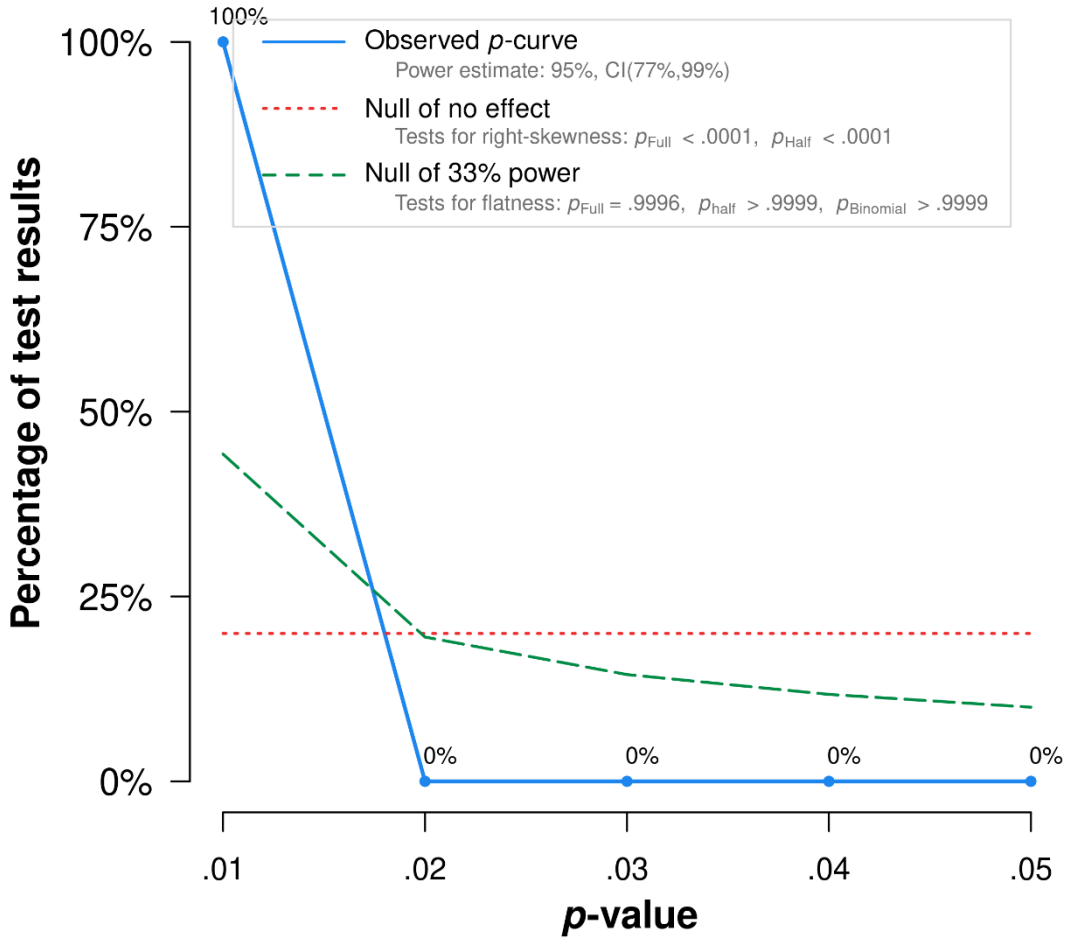
Forms of contact	Women		Men		Total	χ^2	*df*
	N	%	*N*	%			
In person, phone, letter	712	46.5	818	53.5	1530	18.01*	1
Email	529	47.6	582	52.4	1111	7.12**	1
Paper petition	842	54.6	700	45.4	1542	9.40**	1
Online petition	536	58.8	376	41.2	912	24.66*	1

N is the weighted number of cases. Only the number of respondents who answered "yes" was reported in the total

*$p<.001$, **$p<.01$

(Brundidge et al., 2013, p. 11)

In Figure D1, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $\chi^2(1) = 18.01$; $\chi^2(1) = 7.12$; $\chi^2(1) = 9.40$; $\chi^2(1) = 24.66$.



Note: The observed *p*-curve includes 4 statistically significant ($p < .05$) results, of which 4 are $p < .025$.
There were no non-significant results entered.

Figure D1. The single-article *p*-curve for the number 2 of the bottom 10 (i.e., Brundidge et al., 2013).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D2, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D2, the 33% power test is $p = .9996$ for the full *p*-curve, for the half p-curve is $p > .9999$, and for the binomial 33% power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.0625	*Z*=-5.13, *p*<.0001	*Z*=-4.52, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*>.9999	*Z*=3.32, *p*=.9996	*Z*=4.11, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 95% 90% Confidence interval: (77% , 99%)		

Figure D2. Additional statistics for the single-article *p*-curve for the number 2 of the bottom 10 (i.e., Brundidge et al., 2013).

Number 3 of the Bottom 10

Number 4 of the Bottom 10

The number 4 of the bottom 10 is the first study reported in the paper *Social Norms and Egalitarian Values Mitigate Authoritarian Intolerance Toward Sexual Minorities* (Oyamot et al., 2016). Table D4 shows the score on each of the eight selected RDF for the number 4 paper.

Table D4

Coding paper nr. 4 of the bottom 10 (i.e., Oyamot et al., 2016)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "According to our model, one way that authoritarian intolerance could be curtailed by the social context is through clear norms and pressures promoting tolerance. (…) Extending this model to attitudes toward sexual minorities, we predicted that as norms become more tolerant toward this group, so too will authoritarians' attitudes." (Oyamot et al., 2016, p. 781) "our model of authoritarianism-in-social-context suggests that authoritarians' attitudes will shift toward greater tolerance. Extending our theoretical model to the case of attitudes toward sexual minorities, we formulated three main hypotheses, which we tested using relevant data from the 1992, 2000, 2004, 2008, and 2012 American National Election Studies: 1) There will be a subset of authoritarians who endorse humanitarian-egalitarian values (i.e., egalitarian authoritarians). We expect the correlation between authoritarianism and endorsement of egalitarian values will be negative but small and that the number of individuals categorized as traditional authoritarians will be comparable to the number categorized as egalitarian authoritarians. 2) Both traditional authoritarian and egalitarian authoritarian attitudes will become more tolerant between 1992 and 2012. The strongest version of this prediction would be that attitudes would shift from negative to positive over this time frame. However, given the deep antipathy authoritarians have felt toward sexual minorities, we expected that their attitudes would move from negative to "less negative" or neutral. 3) Egalitarian authoritarians will generally hold more positive attitudes toward sexual minorities than traditional authoritarians at each time point between 1992 and 2012." (Oyamot et al., 2016, p. 781-782) Exploratory: "Based on the prior research, it was unclear whether changing norms would have stronger influence over authoritarians' attitudes than those of nonauthoritarians. A secondary goal of this study was to explore relative rates of change (or stability) in authoritarian versus

		nonauthoritarian attitudes toward sexual minorities between 1992 and 2012, as well as comparing the attitudes of egalitarian versus traditional authoritarians." (Oyamot et al., 2016, p. 781) "We regarded as an open question whether magnitude of attitude change would differ between authoritarians and nonauthoritarians and between authoritarian subtypes, and we conducted exploratory analyses regarding these groups' attitudes. Our theory is primarily concerned with examining authoritarians' attitudes, but nonauthoritarians' attitudes provided a useful comparison for understanding the extent of tolerance changes (or lack thereof) during the period examined." (Oyamot et al., 2016, p. 782-783)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	0	"We conducted secondary analyses on all available ANES datasets that included measures of child-rearing values (authoritarianism proxy variable), egalitarianism, and attitudes towards gay men and lesbians (1992, 2000, 2004, 2008, 2012).[2] These datasets are composed of representative samples of the adult U.S. population, and sample size varied between 1212 and 5474 (ANES, 1992–2012)." (Oyamot et al., 2016, p. 783)
Sharing/Openness (i.e., materials, data, code).	1	Data: "American National Election Studies. (199222012). User's guide and codebook for the ANES 1992, 2000, 2004, 2008, 2012 time-series studies. Ann Arbor, MI and Palo Alto, CA: University of Michigan and Stanford University. Retrieved from http://electionstudies.org/studypages/download/datacenter_all_NoData.php" (Oyamot et al., 2016, p. 792) "Using data from the American National Election Studies (ANES) collected between 1992 and 2012" (Oyamot et al., 2016, p. 783)

"using relevant data from the 1992, 2000, 2004, 2008, and 2012 American National Election Studies" (Oyamot et al., 2016, p. 782)

Materials:
"Authoritarian tendencies were scored using four questions about child-rearing values that have been found to be a valid proxy measure for authoritarianism (Brandt & Reyna, 2014; Oyamot, Borgida, et al., 2006, Stenner, 2005). Participants were given a series of four paired qualities and indicated which was more important to foster in a child (independence or respect for elders; curiosity or good manners; self-reliance or obedience; being considerate or well-behaved). Responses to each item were coded such that higher scores reflected authoritarian tendencies: 1 (inconsistent with authoritarian predispositions; the first quality of each pair listed above), 2 (the respondent indicated that both qualities were important), or 3 (consistent with authoritarian predispositions; the second quality of the pairs above). Responses were then averaged to create an authoritarianism score. Scale calculations were conducted separately for each dataset, and the psychometric properties were similar for all datasets (α's = .59 – .68, *M*'s = 2.16 – 2.34, *SD*'s = .55 – .61). The scale reliability in our study was comparable to those in other studies using this scale (e.g., Feldman & Stenner, 1997; Stenner, 2005). Descriptive statistics (scale a, range, mean, SD) for all continuous variables are shown in Table 1.
Endorsement of egalitarian values was measured using the six-item egalitarianism scale ("Our society should do whatever is necessary to make sure that everyone has an equal opportunity to succeed"; "We have gone too far in pushing equal rights in this country"; "One of the big problems in this country is that we don't give everyone an equal chance"; "This country would be better off if we worried less about how equal people are"; "It is not really that big a problem if some people have more of a chance in life than others"; "If people were treated more equally in this country, we would have many fewer problems."), scored on a 5-point Likert scale so that higher scores indicated stronger endorsement of egalitarianism. The psychometric properties of the scale were consistent across the datasets (α's = .66 – .78, *M*'s = 3.41 – 3.57, *SD*'s = .76 – .85).
The ANES uses the specific terms "gay men and

		lesbians" and "gay rights" in its questions. In our research, these terms correspond to sexual minorities and LGBT rights; we chose the latter terms as they are the more inclusive and contemporary terminology. The primary outcome variable was attitudes toward sexual minorities (i.e., gay men and lesbians) as measured by the Feeling Thermometer (FT). Responses on the FT ranged from 0 (*very cold or unfavorable feeling*) to 100 (*very warm or favorable feeling*), with 50 indicating neutral feelings (*no feeling at all*). We also examined respondents' attitudes on LGBT rights issues measured in the ANES datasets." (Oyamot et al., 2016, p. 783)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	The third footnote: "While multiple regression approaches are often used in the analysis of continuous variables, we opted for an ANOVA approach here because it is the most appropriate method for our research questions examining mean differences in attitudes across dataset years, egalitarian-authoritarian combinations, and dataset year by authoritarian subgroups. A multiple regression approach addresses the slightly different question of whether the relationship between contiguous dataset years (predictor) and attitude (DV) differs as a function of value combinations (moderator). However, in the interest of thoroughness, we also conducted multiple regression analyses modeling this three-way interaction effect and found the same pattern of results as shown in the ANOVA. In earlier years where sample sizes are smaller (e.g., 1992), the ANOVA/quartile split approach yields attitude estimates that are larger than the multiple regression approach. However, in general the estimates across datasets and the patterns among egalitarian-authoritarian combinations are very similar." (Oyamot et al., 2016, p. 785)
Fallacious interpretation of (lack of) statistical significance.	0	Effect sizes (Cohen's *d*) are reported. "Results indicated that (1) there was a subset of authoritarians who endorsed egalitarian values, (2) authoritarians in general became more tolerant (i.e., held less negative attitudes) toward sexual minorities between 1992 and 2012, and (3)

		"egalitarian authoritarians" held more positive attitudes toward sexual minorities than other authoritarians. The findings contribute to contemporary theory and research on authoritarianism, which is moving from a monolithic view of authoritarianism to one in which culture and core values activate and shape manifestations of authoritarian tendencies." (Oyamot et al., 2016, p. 777) "Furthermore, the number of individuals who could be categorized as traditional authoritarians was comparable to the number categorized as egalitarian authoritarians. Using quartile splits on the authoritarianism and egalitarianism variables, we categorized participants into one of four authoritarian-egalitarian combinations for each dataset: *traditional authoritarians* (upper 25% Child-Rearing Values (CRV) and lower 25% egalitarianism scores), *egalitarian authoritarians* (upper 25% CRV and upper 25% egalitarianism scores), *egalitarians* (lower 25% CRV and upper 25% egalitarianism scores), *nonegalitarians* (lower 25% CRV and lower 25% egalitarianism scores). Corroborating the low correlations between authoritarianism–egalitarianism, samples sizes, and frequencies for each value combination indicated comparable numbers for the two types of authoritarians (see Table 2). Collapsed across all datasets, traditional authoritarians made up 6.2% of the sample (n = 780) and egalitarian authoritarians 7.6% (n = 953)." (Oyamot et al., 2016, p. 784)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"As in our previous studies, the correlation between authoritarianism and endorsement of egalitarian values, though statistically significant, was only weakly associated across all years represented in the ANES data (r's = -.13 to -.08, p's < .001)." (Oyamot et al., 2016, p. 784) "To explore trends in attitudes towards sexual minorities and how they related to authoritarianism, egalitarianism, and social norms, we conducted a 4 (authoritarian-egalitarian group) x 5 (dataset year) ANOVA. The omnibus test was followed by Scheffé comparisons to test our specific hypotheses. Unless otherwise noted, all comparisons described are significantly different at p < .001." (Oyamot et al., 2016, p. 784-785) "One general expectation, which was a prerequisite for our second hypothesis, was that societal

attitudes towards sexual minorities had become more tolerant between 1992 and 2012, and this was indeed the case. The main effect for dataset year was significant, $F(4, 3986) = 21.86$, $p < .001$. Follow-up Scheffé comparisons between each chronologically contiguous dataset showed that respondents in 2000 expressed more positive attitudes toward sexual minorities ($M = 48.57$, $SD = 28.78$) than respondents in 1992 ($M = 39.20$, $SD = 28.39$), $p < .001$, $d = .32$. There were no other significant differences between contiguous datasets." (Oyamot et al., 2016, p. 785)

"Results supported Hypothesis 2: both authoritarian types were less negative toward the group in 2012 than in 1992. Moreover, we found evidence that the attitude-change trajectories of traditional authoritarians and egalitarian authoritarians diverged. Specifically, the year by authoritarian subgroup interaction was significant, $F(12, 3986) = 3.22$, $p < .001$, indicating that attitudes differed as a function of both authoritarian-egalitarian group and year. To further explore this interaction, we performed a one-way ANOVA for each authoritarian-egalitarian group, which indicated that the main effect of year was significant for each group, F's$(4, 748 – 1436) > 5.56$, p's $<.001$. Consistent with Hypothesis 2, Scheffé comparisons of each contiguous dataset showed that traditional authoritarians became less negative toward sexual minorities between 1992 and 2012, and this change occurred primarily between 1992 and 2000. Traditional authoritarian respondents in 2000 were significantly less negative toward sexual minorities ($M = 33.65$, $SD = 27.81$) than their 1992 counterparts ($M = 17.45$, $SD = 21.86$), $p < .01$, $d = 0.64$. However, between 2000 and 2012, traditional authoritarians' attitudes remained essentially the same, with no significant differences between contiguous datasets. Therefore, as predicted by our model, authoritarians appeared to be responsive to changing societal norms, although it is important to note the limits of this responsiveness: Traditional authoritarians' attitude shift cannot be precisely characterized as increasing tolerance as they had only shifted from "quite cold" to merely "fairly cold" on the feeling thermometer measure.
As expected, egalitarian authoritarians showed even stronger trends toward tolerance of sexual minorities between 1992 and 2012. Like traditional authoritarians, egalitarian-authoritarian respondents

in 2000 held less negative attitudes toward sexual minorities (M = 45.80, SD = 28.92) than their 1992 counterparts (M = 31.10, SD = 28.38), $p < .01$, d = 50.52. Similar to traditional authoritarians, there was a period of quiescence in egalitarian-authoritarian attitudes in which there were no significant attitudinal shifts between 2000 and 2008. The magnitude of the 1992–2000 attitude shift was similar for both authoritarian types. Each showed a 15–16 degree change in mean attitude, and the effect sizes were of a similar magnitude (.64 and .52).

Unlike traditional authoritarians, egalitarian authoritarians showed another significant increase in tolerance between 2008 (M = 42.93, SD = 29.69) and 2012 (M = 52.68, SD = 27.98), $p < .01$, d = 0.35. Egalitarian authoritarians appeared to be more responsive to shifting norms of tolerance, having moved from "fairly cold" to "no feeling at all" (or neutral) toward sexual minorities, with these shifts coinciding with periods of rapid change in social norms.

Egalitarians and nonegalitarians showed a similar pattern of increasing tolerance as traditional authoritarians, though their initial 1992 attitudes were somewhat more positive (M = 59 and 34, respectively): significant increases between 1992 and 2000 (p's < .05, d = .64 and .73, respectively), followed by no significant changes through 2012.

Our third hypothesis related to egalitarian authoritarians having more positive attitudes toward sexual minorities than traditional authoritarians at each time point, and this was also supported. *H3*: Egalitarian authoritarians will generally have more positive attitudes toward sexual minorities than traditional authoritarians. The main effect for authoritarian-egalitarian categorization was significant, $F(3, 3986) = 227.23$, $p < .001$. Follow-up comparisons showed that, collapsed across all years, egalitarian authoritarians had more positive attitudes toward sexual minorities (M = 47.39, SD = 29.18) than traditional authoritarians (M = 33.10, SD = 27.91), d = .50. Significant differences were also found within each individual year, except for 2008." (Oyamot et al., 2016, p. 785-786)

"We next examined opinions on LGBT rights issues in the 1992, 2000, 2008, and 2012 ANES datasets and found trends similar to those for attitudes toward sexual minorities. As with group attitudes, and consistent with Hypothesis 2, support for

> protection from workplace discrimination generally increased between 1992 and 2012 (see Figure 3). The two-way ANOVA showed significant main effects for year and authoritarian subgroup, *F*'s(3, 2356) > 20, *p*'s < .001, qualified by a significant interaction, *F*(9, 2356) = 3.03, *p* < .01. Scheffé post hoc tests between years within each authoritarian-egalitarian group showed that traditional authoritarians moved from opposition to job protections for sexual minorities in 1992 (*M* = 2.2, *SD* = 1.5) to a neutral position by 2008 (*M* = 3.2, *SD* = 1.6), *p* < .01, *d* = 0.64, which then remained unchanged through 2012. Egalitarian authoritarians moved from neutral (*M* = 3.5, *SD* = 1.7) in 1992 to moderate support in 2012 (*M* = 4.1, *SD* = 1.5), *p* < .05, *d* = 0.38. Consistent with Hypothesis 3, egalitarian authoritarians' attitudes were significantly more supportive of sexual minorities' job-protection rights than traditional authoritarians' opinions between 1992 and 2012. Egalitarians strongly supported job protections, and their stance was stable across the time period, (*M* = 4.5 – 4.7, *SD* = .85 – .97). Nonegalitarians showed significant increases in support on this issue between 2000 (*M* = 3.1, *SD* = 1.6) and 2008 (*M* = 3.9, *SD* = 1.5), *p* < .03, *d* = 0.52.
>
> *Military service.* Changes in support for sexual minorities serving in the military followed essentially the same pattern as those for job protection. The two-way ANOVA showed significant main effects for year and authoritarian subgroup, *F*'s(3, 2394)> 41, *p*'s < .001, qualified by a significant interaction, *F*(9, 2394) = 4.49, *p* < .001. Each authoritarian-egalitarian combination group showed significantly more tolerant opinions between 1992 and 2000 (see Figure 4) *p*'s < .01, *d*'s > 0.45. In addition, traditional authoritarians showed another significant change between 2000 and 2008, *p* < .05, *d* = 0.41, but no change in 2012. In contrast, egalitarian authoritarians' attitudes became more supportive of sexual minorities in the military between 2000 and 2012, *p* < .05, *d* = 0.33." (Oyamot et al., 2016, p. 786-787)

In Figure D3, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): *F*(4, 3986) = 21.86; *F*(12, 3986) = 3.22; *F*(3, 3986) = 227.23; *F*(9, 2356) = 3.03; *F*(9, 2394) = 4.49.

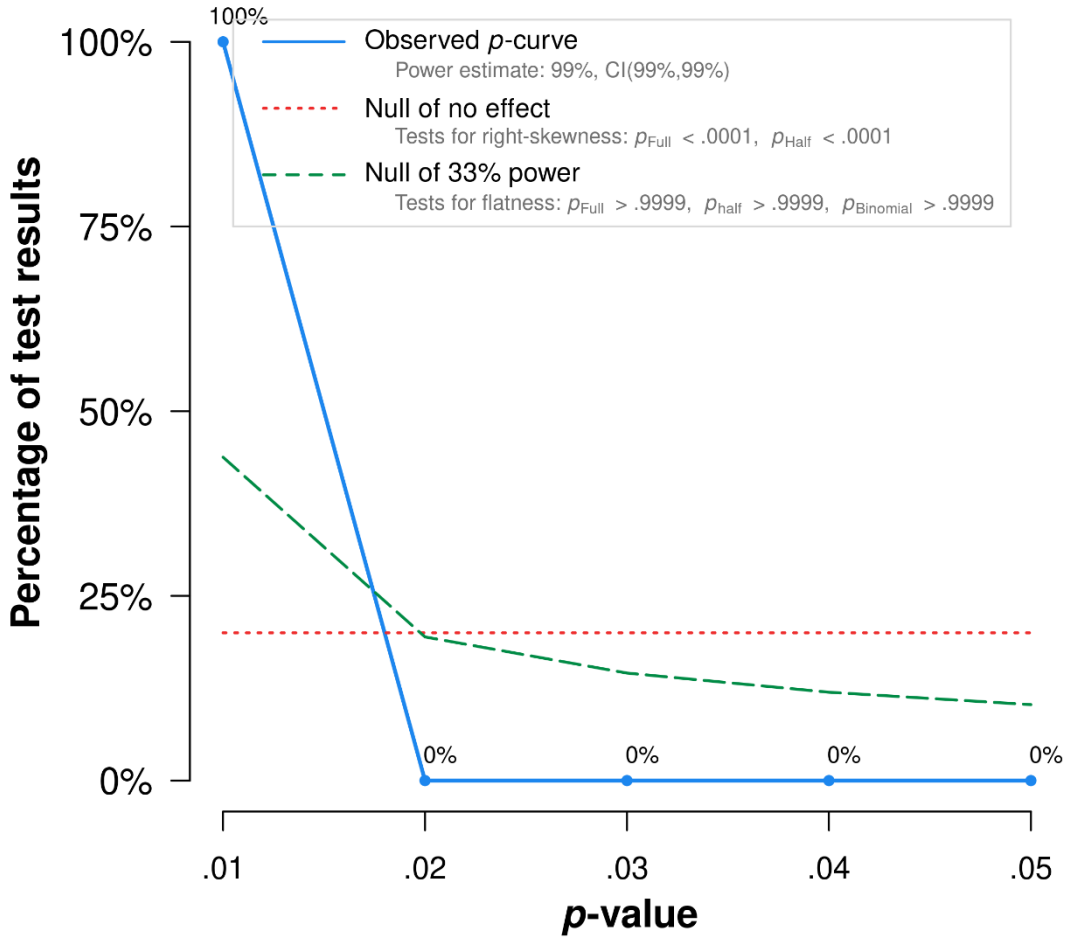


Figure D3. The single-article *p*-curve for the number 4 of the bottom 10 (i.e., Oyamot et al., 2016).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D4, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D4, the 33% power test is $p > .9999$ for the full *p*-curve, for the half p-curve is $p > .9999$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.0313	Z=-10.68, p<.0001	Z=-10.27, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p>.9999	Z=7.81, p>.9999	Z=8.36, p>.9999
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure D4. Additional statistics for single-article p-curve for the number 4 of the bottom 10 (i.e., Oyamot et al., 2016).

Number 5 of the Bottom 10

The number 5 of the bottom 10 is the first study reported in the paper *Personality profiles in substance use disorders: Do they differ in clinical symptomatology, personality disorders and coping?* (Santens et al., 2018). Table D5 shows the score on each of the eight selected RDF for the number 5 paper.

Table D5

Coding paper nr. 5 of the bottom 10 (i.e., Santens et al., 2018)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "This study aimed to expand the existing literature on BIS/BAS, EC and SUDs. Within a large, clinical sample of Caucasian patients, we explored whether it is possible to establish subgroups of patients based on temperamental factors. We expected to find two clusters of personality profiles: an impulsive/disinhibited group with high reward-sensitivity (high BAS, low BIS) and an anxious/inhibited group (low BAS, high BIS). In both groups we expected to find rather low levels of effortful control, as it is assumed that a high level of self-control (EC) is a protective factor in developing psychopathology (Nigg, 2006; Rothbart & Sheese, 2007). We also explored if the clusters we identified differed in clinical symptomatology, personality disorders and coping styles. As some

		research suggest a relationship between type of substance use and temperamental/personality factors (e.g. the traits "novelty seeking or sensation seeking" are associated with experimentation and abuse of several substances), whereas indicators of poor self-regulation correspond to the gradient of substance use categories (Conway, Kane, Ball, Poling, & Rounsaville, 2003), we also explored whether there are differences in terms of substance used in the clusters." (Santens et al., 2018, p. 62)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	2	"Twelve patients were excluded on the basis of multivariate outliers prior to conducting cluster analysis, resulting in a final sample of 700 patients (68.1% males and 31.9% females)." (Santens et al., 2018, p. 62)
Sample size (predetermined or not).	3	"The study included 712 consecutive admitted adult Caucasian patients on a specialized, inpatient treatment program for SUDs. Twelve patients were excluded on the basis of multivariate outliers prior to conducting cluster analysis, resulting in a final sample of 700 patients (68.1% males and 31.9% females). Diagnosis of SUD (dependence or abuse) based on DSM-IV-TR criteria (APA, 2000) was made by experienced psychiatrists (ES HP. The mean age of the participants was 45.7 years (SD=11.25). The computerized self-report questionnaires were administered during the second week of admission (after detoxification) on the addiction ward. All patients signed an informed consent paper and the research was approved by the ethics committee of the hospital." (Santens et al., 2018, p. 62) "Computerized self-report questionnaires were administered to 712 adult patients admitted to a specialized inpatient treatment program for SUDs." (Santens et al., 2018, p. 61)
Sharing/Openness (i.e., materials, data, code).	2	Materials: "The Behavioral Inhibition/Behavioral Activation System Scales (BIS/BAS; Carver & White, 1994) is a self-report questionnaire that consists of 24 items which are rated on a 4-point Likert scale (ranging from 1 = *I totally agree* to 4 = *I totally disagree*). It

measures the reactivity of two motivational systems. The BIS responds to cues associated with punishment and non-reward while the BAS reflects sensitivity to reward.
The BIS and BAS total scales demonstrated acceptable internal consistency coefficients in the present sample (α = 0.76 and 0.85 respectively)." (Santens et al., 2018, p. 62)

"The 19-item Effortful Control (EC) Scale of the Adult Temperament Questionnaire Short-Form (Rothbart, Ahadi, & Evans, 2000) was used to measure self-regulatory capacity. Participants rated their general capacity to exert behavioral and attentional control on a 7-point Likert scale. The EC total score demonstrated acceptable internal consistency in the present sample (α =0.80)." (Santens et al., 2018, p. 62)

"The Symptom checklist-90-Revised (SCL-90-R, Arindell & Ettema, 2003, Dutch version), is a questionnaire that assesses severity of psychological symptoms of depression (DEP), anxiety (ANX), agoraphobia (AGO), somatization (SOM), insufficiency of thought and behaviour (IN), hostility (HOS), sleeping problems (SLE), distrust and interpersonal sensitivity (DIS). Patients are asked to rate the 90 items on a five-point Likert scale. The internal consistency, test-retest reliability and convergent validity of this measure in adult psychiatric outpatients is supported by previous research (Arindell, Boonsma, Ettema, & Stewart, 2004).
In the present study the Cronbach's alphas are the following: DEP = 0.93, ANX = 0.91, AGO = 0.85, SOM = 0.84, IN = 0.91, HOS = 0.77, SLE = 0.80, DIS = 0.87, representing acceptable internal consistency." (Santens et al., 2018, p. 62)

"The Assessment of DSM-IV Personality Disorders (ADP-IV, Schotte & De Doncker, 1996), a 94-item Dutch self-report questionnaire, assesses the PDs criteria of the 10 personality disorders, described in the DSMIV-TR (American Psychiatric Association, 2000). Items on the ADP-IV are first rated on the typicality of the trait to the respondent (1 = *totally not*, 7 = *totally true*). For items that are rated as applicable at a moderate or higher level (trait score > 5), the participant also has to rate the distress for the participant or his/her environment on a 3-point Distress scale (1 = *totally not*, 3 = *almost always*). The dimensional scale scores demonstrated

		marginally acceptable to acceptable internal consistency coefficients in the present sample (ranging from α = 0.68 to α = 0.87).” (Santens et al., 2018, p. 62)
		“The Utrecht Coping List (UCL, Schreurs, van de Willige, Brosschot, Telligen, & Graus, 1993), a self-report questionnaire with 47 items, assesses how people usually react when confronted with stressful situations. The UCL has been found to have satisfactory psychometric properties in a Dutch population. There are 7 scales to distinguish different coping styles namely active coping (ACT), avoidant coping (AVOI), passive coping (PAS), seeking social support (SOC), reassuring thoughts (REA), expression of emotions (EXP), palliative coping (PAL). The participants must rate their answers on a 4 point Likert scale. In the present sample the Cronbach's alphas are the following: ACT = 0.86, AVOI = 0.73, PAS = 0.80, SOC = 0.86, REA = 0.64, EXP = 0.62, PAL = 0.68.” (Santens et al., 2018, p. 62)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	“To identify subtypes of individuals receiving inpatient treatment for SUD based on reactive and regulative temperament dimensions, we performed a two-step cluster analysis on the standardized BIS/BAS and EC scale scores (i.e., *z*-scores). Cluster analysis aims to group patients into relatively homogeneous clusters in such a way that patients within one cluster have more in common than they do with patients assigned to other clusters (Gore Jr., 2000). First, a hierarchical cluster analysis was carried out using Ward's method based on squared Euclidian distances. Second, these initial cluster centers were subsequently used as non-random starting points in a k-means clustering procedure (MacQueen, 1967), resulting in an optimized cluster solution. To validate the clusters, we made use of the multivariate analysis of variance (MANOVAs) with the SUDs subtypes as independent variable and clinical symptomatology, personality disorders, and coping as dependent variables. A chi square analysis was performed in order to ascertain whether SUDsubtypes differ in

		terms of type of substance used." (Santens et al., 2018, p. 63)
Fallacious interpretation of (lack of) statistical significance.	2	"A chi square analysis showed a significant association between type of substance used and cluster membership, $\chi^2(6) = 55.87$, $p < 0.001$. Patients who only abuse alcohol are most prevalent in the Resilient cluster, those who abuse alcohol and benzodiazepines are most frequently found in the Resilient and Anxious cluster. Patients who abuse alcohol and drugs are found most frequently in the Anxious and Reward-Sensitive cluster" (Santens et al., 2018, p. 63) "The Anxious cluster showed the highest scores on each of the clinical symptoms. The Resilient cluster consistently displayed the lowest scores on all clinical symptoms (except for agoraphobia). The Reward-Sensitive cluster had overall in between scores" (Santens et al., 2018, p. 63) "The Resilient cluster scored significantly higher on 5 of the 7 coping styles except on active coping (same score as the Reward-Sensitive cluster) and reassuring thoughts (same score as the Anxious cluster). The Anxious cluster, scored higher on the passive and avoidant coping style and the lowest on the active coping style. The Reward-Sensitive cluster, had the highest scores on expression of emotions and reassuring thoughts as coping style. Patients of the Anxious and the Reward-Sensitive cluster scored significantly higher on palliative coping style and social support seeking compared to patients of the Resilient cluster" (Santens et al., 2018, p. 63-64) "The Resilient cluster consistently displayed the lowest scores on Cluster A, B and C PDs except for the schizoid and avoidant personality disorder where they did not differ from the Reward-Sensitive cluster. Consistent with our expectations, we found the highest scores in the Anxious cluster on Cluster C PD pathology (avoidant, dependent and obsessive compulsive personality disorder) and on the Borderline and Histrionic personality disorder for the Cluster B pathology. In the Reward-Sensitive cluster we found as expected the highest scores on the antisocial and narcissistic traits (Mowlaie, Abolghasemi, & Aghababaei, 2016); although they were not significant different

		in comparison with the Anxious cluster" (Santens et al., 2018, p. 64)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"We found more women than men in the Anxious cluster, whereas more men than women were present in the Resilient and the Reward-Sensitive cluster ($\chi^2_{(2)} = 35.34$, $p < 0.001$). The results of an ANOVA showed a statistically significant difference in age between the three clusters ($F(2, 697) = 24.82$; $p < 0.001$). Post-hoc comparisons learned that the patients in the Resilient cluster ($M = 49.27$, $SD = 10.10$) were significant older than the patients in the Anxious ($M = 44.19$, $SD = 11.43$) and the Reward-Sensitive cluster ($M = 42.63$, $SD = 11.26$). Mean age between the Anxious and the Reward-Sensitive cluster did not significantly differ." (Santens et al., 2018, p. 63) "A chi square analysis showed a significant association between type of substance used and cluster membership, $\chi^2(6) = 55.87$, $p < 0.001$." (Santens et al., 2018, p. 63) "The MANOVA comparing the three clusters (independent variable) on clinical symptoms as assessed by means of the SCL-90 (dependent variable) showed significant differences (Wilks *Lambda* = 0.81, $F(18, 1378) = 8,75$, $p = 0.000$) between the three clusters. The follow-up univariate analysis showed differences on all domains." (Santens et al., 2018, p. 63) "The results of the MANOVA with the three clusters as independent variable and the UCL scales as dependent variables showed overall significant differences between the three clusters (Wilks *Lambda* = 0.72, $F(14, 1382) = 17.62$, $p = 0.000$)." (Santens et al., 2018, p. 63) "The MANOVAs comparing the three clusters on Axis-II pathology assessed by the ADP-IV revealed significant overall differences (Wilks' *Lambda*: 0.057, $F(24, 1368) = 18.58$, $p = 0.000$)." (Santens et al., 2018, p. 64)

In Figure D5, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $\chi^2(2) = 35.34$; $F(2, 697) = 24.82$; $\chi^2(6) = 55.87$; $F(18, 1378) = 8.75$; $F(14, 1382) = 17.62$; $F(24, 1368) = 18.58$.

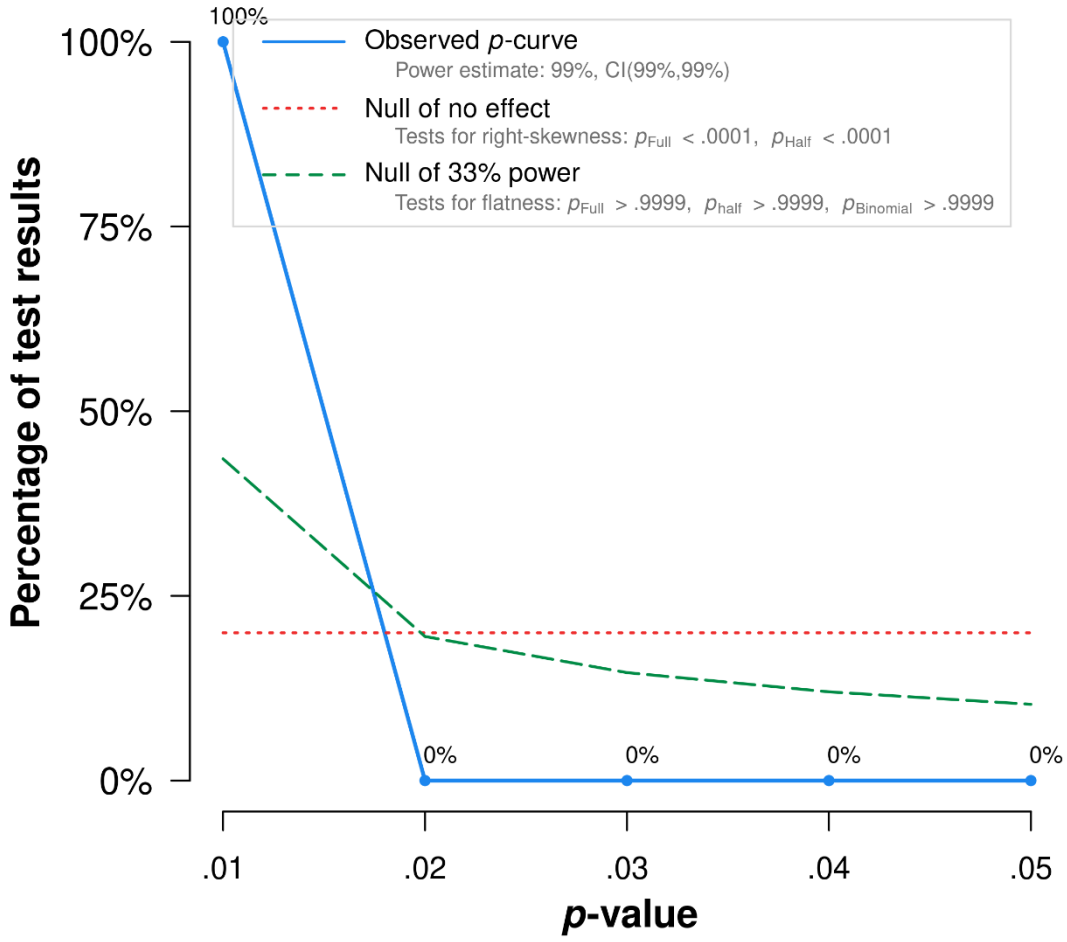


Figure D5. The single-article *p*-curve for the number 5 of the bottom 10 (i.e., Santens et al., 2018).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D6, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D6, the 33% power test is $p > .9999$ for the full *p*-curve, for the half p-curve is $p > .9999$, and for the binomial 33%

power test is $p > .9999$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test (Share of results p<.025)	Continuous Test (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	Half p-curve (p's<.025)
1) Studies contain evidential value. (Right skew)	p=.0156	Z=-16.3, p<.0001	Z=-16.04, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p>.9999	Z=12.57, p>.9999	Z=12.92, p>.9999
	Statistical Power		
Power of tests included in *p*-curve (correcting for selective reporting)	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure D6. Additional statistics for single-article *p*-curve for the number 5 of the bottom 10 (i.e., Santens et al., 2018).

Number 6 of the Bottom 10

The number 6 of the bottom 10 is the first study reported in the paper *Pretrial Predictors of Judgments in the O. J. Simpson Case* (Peacock et al., 1997). Table D6 shows the score on each of the eight selected RDF for the number 6 paper.

Table D6

Coding paper nr. 6 of the bottom 10 (i.e., Peacock et al., 1997)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Exploratory: "we examined a number of case-relevant and general attitudinal variables to explore consistency in beliefs and attitudes prior to the trial, and thus, prior to established facts." (Peacock et al., 1997, p. 442) "This study's purpose was to examine pretrial correlates of attitudes regarding Simpson's guilt. In retrospect, we were able to examine whether the issues of racism and domestic violence that emerged after the trial were important in judgments of guilt before the trial. We examined predictors of community and four-year college students' judgments of Simpson's guilt or innocence two to three months before the trial began." (Peacock et al., 1997, p. 441) "a further purpose of our study was to examine ethnicity, namely African Americans versus non-African Americans. Though public opinion polls

		before, during, and after the trial consistently reported that African Americans were more likely to believe Simpson innocent than were Caucasians, we were interested in whether the same predictors would explain perceived innocence or guilt for both African Americans and non-African Americans. That is, is ethnicity a moderator, such that different variables are useful in explaining African Americans' and non-African Americans' judgments of guilt? Alternatively, would the same predictors explain perceptions of guilt for both groups?" (Peacock et al., 1997, p. 442) "In the present study, our purpose was not to focus on ethnic difference but to examine whether the same predictors explained both African Americans' and non-African Americans' beliefs about Simpson's guilt. Another question concerned the moderation of innocence or guilt predictors by the salience of the issue. By salience, we mean the extent of attention to the case" (Peacock et al., 1997, p. 442) "suggest that we would find stronger relations between predictors and judgments of guilt among those for whom the case was more salient." (Peacock et al., 1997, p. 443) "the current study examined proximal beliefs about the case, more general attitudes and beliefs, potentially relevant experiences, and respondent attributes as predictors of the belief that Simpson was guilty." (Peacock et al., 1997, p. 443-444)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	"The participants were 578 community college (40%) and four-year state university students (60%). The average age of the sample was 26.8 (*SD* = 9.26), with no significant age difference between African Americans and non-African Americans. The ethnic distribution was 15.7% (90) African

		Americans, 3.8% (22) Asian Americans, 57.4% (329) Caucasians, 18.2% (104) Hispanics, and 4.9% (28) "Other." Of those identifying their gender, 144 were men and 432 women. Participants' annual median family income was $30,000 to $40,000. Data were collected from October 1, 1994, to November 30, 1994, two months prior to the start of the trial, and Table 1 summarizes the events known to the participants at the time of the survey." (Peacock et al., 1997, p. 444) "Five hundred seventy-eight community college and four-year state university students responded to questionnaires designed to assess judgments regarding O. J. Simpson's guilt, beliefs surrounding the case, general attitudes, and background information." (Peacock et al., 1997, p. 441)
Sharing/Openness (i.e., materials, data, code).	2	Materials: "Participants were recruited through classroom visitations and through the normal departmental policy of posting ongoing experiments. Students were told the study examined the diversity of attitudes and beliefs surrounding the Simpson case. Participants completed questionnaires at home and returned them to instructors or to designated offices. The nine-page questionnaire took approximately thirty minutes to complete. The questionnaire consisted of items assessing participants' judgments regarding the upcoming trial and Simpson's guilt, beliefs surrounding the case, general attitudes, and background information. Cronbach's alphas were used to refine all scales to produce the most reliable measures. The dependent variable item, first on the questionnaire, was, "Regarding the murder of Nicole Brown Simpson and Ron Goldman, O. J. Simpson is. . ." The response alternatives ranged from 1 (*definitely innocent*) to 7 (*definitely guilty*)." (Peacock et al., 1997, p. 445) "*Case Relevant Beliefs*. Thirty-one items were designed to ask about issues relevant at the time of the survey and used seven-point scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*)." (Peacock et al., 1997, p. 445) "*Salience*. An eight-item scale, scored in the direction of high salience, assessed the case's salience by asking respondents on a five-point scale

		from 1 (*not at all* or *never*) to 5 (*all of it* or *frequently*)" (Peacock et al., 1997, p. 445-446) "*Experience Items*. To assess experience with domestic violence, respondents were asked on four-point scales ranging from *never* (1) to *frequently* (4)" (Peacock et al., 1997, p. 446) "*Attitude Measures.* (…) three measures were scored in the direction of violence-supportive beliefs and stereotypes on a seven-point scale, 1 = *strongly disagree* to 7 = *strongly agree*" (Peacock et al., 1997, p. 446) "*Demographic and Background Information*. Participants indicated their gender, ethnicity, age, family, income, and whether they attended a community college or four-year college. They also rated their political orientation on a scale ranging from 1 (*extremely conservative*) to 7 (*extremely liberal*). Participants also indicated their political affiliation, choosing among Democrat, Republican, Independent, and Other." (Peacock et al., 1997, p. 447)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"To determine the predictive utility of beliefs and attitudes in terms of their correspondence to the judgment of Simpson's guilt or innocence, we examined four levels of predictors." (Peacock et al., 1997, p. 443) "Univariate analysis of variance (ANOVA) was performed to determine ethnic differences in the judgment of guilt." (Peacock et al., 1997, p. 447) "The above regression analysis was repeated to assess the moderating effects of salience. High and low salience (median split = 24) was dummy coded and entered into the regression equation such that the interaction variables (e.g., ethnicity x salience, age x salience) were assessed at each step. No significant enhancements of predictability due to the interactions were found." (Peacock et al., 1997, p. 451) "Because ethnicity of respondents (African Americans versus non-African American) was an important consideration in this study, it seemed important to test the incremental predictive utility

		of ethnicity. Two hierarchical analyses were performed with ethnicity entered first and last in the regression equation. Ethnicity accounted for approximately 10% of the variance when entered first but for less than 1% when entered last, a large drop in the proportion accounted for." (Peacock et al., 1997, p. 452)
Fallacious interpretation of (lack of) statistical significance.	2	"To examine gender differences between African Americans and non-African Americans on the dependent variable, a two-way analysis of variance was conducted. No significant gender difference or significant gender/ethnicity interaction was found. Table 2 presents a comparison of selected group means by ethnicity. The means indicated that African Americans were more accepting of interpersonal violence in intimate relationships than were non-African Americans. Non-African Americans were more likely to believe that the media favored Simpson, whereas African Americans were more likely to believe that race was a factor in this case. Mean differences between African Americans and non-African Americans indicated that African Americans were more likely to view Simpson as a role model, and to believe Nicole Brown Simpson was not abused by O. J. Simpson but that the criminal justice system was biased against Blacks. Also, it appears the case was more salient for African Americans than for non-African Americans." (Peacock et al., 1997, p. 447-448) "Despite the significant minority of participants who had experienced domestic violence and the majority who knew domestic violence victims, these variables proved unrelated to Simpson's perceived guilt. Examination of intercorrelations of general attitudes and judgments (the third level) of Simpson's guilt indicated that among African Americans, only public trust was significantly related to judgments of Simpson's guilt. African Americans who indicated more public trust judged Simpson more likely guilty. Among non-African Americans, violence-supportive attitudes were negatively associated with guilt; these correlations, however, were not large." (Peacock et al., 1997, p. 449-450) "A set-wise hierarchical regression analysis was performed to determine the incremental proportion of variance in guilt/innocence judgments attributable to each set. Only those predictors that

		were significant bivariate correlates of perceived guilt in the combined sample were entered into the equation." (Peacock et al., 1997, p. 450)
		"on the basis of our study, the racial polarization emphasized in public polls does not reflect the actuality that greater diversity of views existed within both African American and European American populations than the media would have us believe. African Americans were more likely to perceive the system as unjust and to minimize Simpson's battering than were non-African Americans, thus accounting for the ethnic difference in perceived guilt. However, the differences in perceived guilt did not result simply from ethnicity or race. With all the predictors in the regression equation, ethnicity continued to be individually significant but accounted for little variance. Rather, distrust of the criminal justice system and the belief that Simpson battered his ex-wife helped explain beliefs going into the trial. For each group, these two major predictors accounted for a significant amount of variance. Thus, we did not find uniformity and consensus among African Americans or among non-African Americans." (Peacock et al., 1997, p. 452-453)
		"Salience did not play a strong role in this case. Although balance theory implies that a larger number of predictors and the strength of the combined predictors should characterize the beliefs of those for whom the case was salient more than those for whom the case was not, salience moderated none of the effects. Perhaps the Simpson case was salient to everyone, especially in the Los Angeles-basin area. Another consideration is that the first item on the questionnaire ("Regarding the murder of Nicole Brown Simpson and Ron Goldman, 0. J. Simpson is . . .") activated a cue for a particular attitude, therefore rendering the case salient for each respondent. In conclusion, this pretrial study does not provide sufficient information about what might underlie the specific beliefs about battering and the system in the Simpson criminal case; however, the study does indicate that sentiments surrounding the Simpson verdict are far too complex to attribute to ethnic differences alone." (Peacock et al., 1997, p. 453)
Assessing the evidential value of a single article by	0	"Univariate analysis of variance (ANOVA) was performed to determine ethnic differences in the judgment of guilt. A significant difference was

judging the single-article *p*-curve (Simonsohn et al., 2014).

found for ethnicity, $F(4, 568) = 10.41$, $p < .0005$. Post hoc Tukey's HSD revealed that African Americans ($M = 3.69$) were more likely to perceive Simpson innocent than were Caucasians, Asians, Hispanics, or Other groups, who did not differ, significantly, from each other ($M = 5.12$, 4.95, 4.93, and 4.50, respectively). Therefore, for the remaining analyses, ethnicity was dichotomized (African American, non-African American)." (Peacock et al., 1997, p. 447)

"Also among non-African Americans, political liberalism was negatively related to beliefs that Simpson was guilty. Self-identified Republicans ($M = 5.23$) rated Simpson as more likely guilty than participants identified as Democrats, Independents, or Other, $F(3, 469) = 4.88$, $p < .002$ ($M = 4.65$, 4.64, and 4.55, respectively). Although African American Republicans tended to rate Simpson higher on the guilt scale, only four African Americans identified themselves as Republicans, compared to 173 Republicans among members of other ethnic groups.

The fourth level of analysis focused on the relations between proximal attitudes and Simpson's guilt. Both African Americans and non-African Americans who believed the system was biased against Simpson were less likely to judge him guilty, whereas those who believed Simpson abused Nicole Brown Simpson were more likely to judge him guilty. For these two strongest correlates of perceived guilt, stronger relations were found among African Americans than among non-African Americans (Fisher's $Z = 2.05$, $p < .05$ for system; $Z = 2.11$, $p < .05$ for abuse)." (Peacock et al., 1997, p. 450)

"Finally, the addition of the fourth set, case-related attitudes and beliefs, showed that predictors of guilt/innocence were beliefs pertaining to whether the media favored Simpson, whether Simpson abused Nicole Brown Simpson, and whether the criminal justice system was biased. The $R^2 = .44$ indicated a significant increment of variance (26%) not accounted for by the other sets. In sum, with all the variables in the equation, the strongest predictors were the proximal beliefs that Simpson battered Nicole Brown Simpson and that the justice system was biased." (Peacock et al., 1997, p. 451)

In Figure D7, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(4, 568) = 10.41$; $F(3, 469) = 4.88$; $Z = 2.11$.

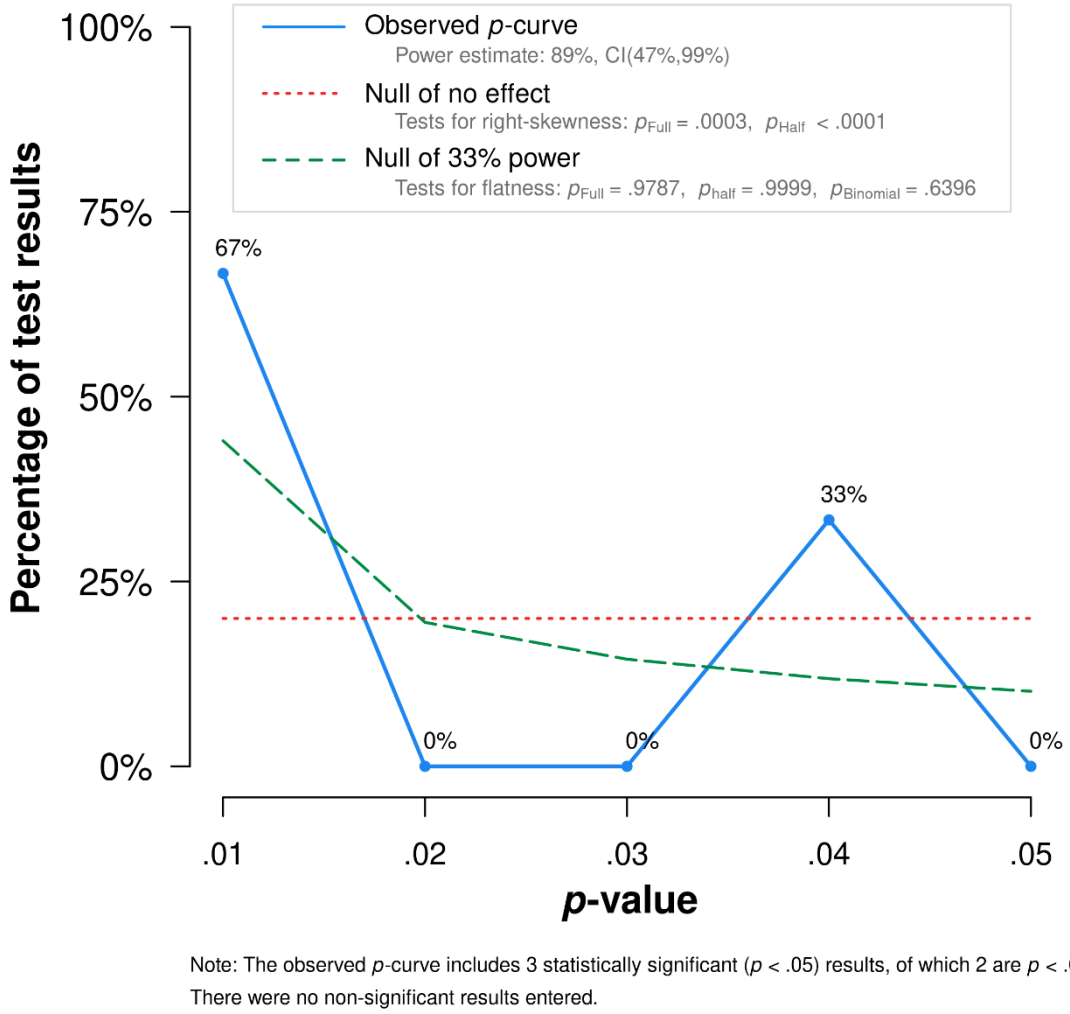


Figure D7. The single-article *p*-curve for the number 6 of the bottom 10 (i.e., Peacock et al., 1997).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D8, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p = .0003$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power

test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D8, the 33% power test is $p = .9787$ for the full *p*-curve, for the half p-curve is $p = .9999$, and for the binomial 33% power test is $p = .6396$; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	p=.5	Z=-3.44, p=.0003	Z=-4.23, p<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p=.6396	Z=2.03, p=.9787	Z=3.66, p=.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 89% 90% Confidence interval: (47% , 99%)		

Figure D8. Additional statistics for single-article *p*-curve for the number 6 of the bottom 10 (i.e., Peacock et al., 1997).

Number 7 of the Bottom 10

The number 7 of the bottom 10 is the first study reported in the paper *Cultural Factors, Depressive and Somatic Symptoms Among Chinese American and European American College Students* (Kalibatseva & Leong, 2018). Table D7 shows the score on each of the eight selected RDF for the number 7 paper.

Table D7

Coding paper nr. 7 of the bottom 10 (i.e., Kalibatseva & Leong, 2018)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "To summarize, the current study examines the relationship between self-construal, loss of face, and emotion regulation, and depressive and somatic symptoms among Chinese American and European American students. It seeks to make a contribution to the existing literature in three ways: (a) provide an empirical test of the relationship between depressive and somatic symptoms among Chinese American and European American college students; (b) examine depressive symptoms, somatic symptoms, self-construal, loss of face, and emotion regulation using a comparative framework; and (c) provide a bridge between group comparisons based

on demographic variables and comparisons based on culturally relevant psychological variables (Helms et al., 2005). Thus, the study poses the following hypotheses:
Hypothesis 1: Chinese American students will somatize by reporting more physical symptoms on the PHQ-15 and the Center for Epidemiologic Studies–Depression (CES-D) somatic subscale than European American students.
Hypothesis 2: Interdependent self-construal, loss of face, and expressive suppression will be positively associated with depressive and somatic symptoms and independent self-construal and cognitive reappraisal will be negatively associated across both groups.
Hypothesis 3: Self-construal, loss of face, and emotion regulation will predict depressive symptoms among Chinese American and European American students above and beyond ethnicity as a predictor." (Kalibatseva & Leong, 2018, p. 1559-1560)

"previous studies of U.S. college students showed a positive association between interdependence and depression and a negative association between independence and depression (Norasakkunkit & Kalick, 2002; Okazaki, 1997, 2000, 2002). Thus, it is hypothesized that in the United States, where independence is valued, independent self-construal serves as a protective factor against depression. Conversely, it is hypothesized interdependent self-construal serves as a risk factor for depression in an individualistic society, along with other culturally salient constructs, such as loss of face." (Kalibatseva & Leong, 2018, p. 1558)

"Given the paucity of empirical studies that test directly the relationship between emotion regulation and somatic and depressive symptoms, the current study tries to fill this gap. Based on the previous research in the United States, it is hypothesized that ES will be positively associated and CR will be negatively associated with depressive and somatic symptoms." (Kalibatseva & Leong, 2018, p. 1558)

"This study focused on loss of face as it has been theorized as an important relational construct, which may negatively affect well-being and help-seeking (Leong, Kim, & Gupta, 2011; Zane & Yeh, 2002). Loss of face (LOF) has a positive association with depressive symptoms and general

		psychological distress among both Asian Americans and European Americans (Leong, Byrne, Hardin, Zhang, & Chong, 2018). Moreover, losing face may be associated with lower levels of seeking mental health services (Cheang & Davis, 2014). Thus, it is hypothesized that an elevated level of concern with losing face has a positive association with somatic and depressive symptoms." (Kalibatseva & Leong, 2018, p. 1559) "This study seeks to fill a gap in the existing empirical literature about the relationship between somatic and depressive symptoms and their associations with cultural factors among Chinese American and European American college students. In particular, the study examined how three culturally relevant psychological constructs, self-construal, loss of face, and emotion regulation, associate with depressive and somatic symptoms among Chinese American and European American college students and if they can explain possible group differences in depressive symptoms." (Kalibatseva & Leong, 2018, p. 1556) "Informed by Katon et al.'s (1982) model and the integration of cross-cultural research methods in racial/ethnic minority research (Leong, Leung, & Cheung, 2010), this study builds on the existing literature by moving from group comparisons based on race and ethnicity to incorporating relevant psychological factors that may explain proposed racial and ethnic differences in self-reported symptoms of depression (Betancourt & Lopez, 1993; Helms, Jernigan, & Mascher, 2005; Leong, Park, & Kalibatseva, 2013). Thus, this study identifies and examines three culturally relevant psychological factors that may be related to depressive and somatic symptoms: self-construal, loss of face, and emotion regulation." (Kalibatseva & Leong, 2018, p. 1556)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to	0	No exclusions.

deal with outliers in an ad hoc manner.		
Sample size (predetermined or not).	3	"The sample consisted of 519 participants predominantly from two large Midwestern universities. There were 204 (39.3%) participants who self-identified as Chinese American. Almost two thirds of the Chinese American sample (64.2%, *n* = 131) were female and 35.8% (n = 73) were male. The mean age for the Chinese American sample was 20.65 (*SD* = 2.95). There were 315 participants (60.7%) who self-identified as European American. Sixty-two percent (*n* = 196) identified as female and 38% (*n* = 120) as male. The mean age was 19.87 (*SD* = 2.88)." (Kalibatseva & Leong, 2018, p. 1560) "Participants were recruited through the university participant pool, targeted emails from the Registrar's Office, campus organizations of Asian American students, and a posting on the list-serv of the Asian American Psychological Association. To facilitate the recruitment of Chinese American students at one of the universities, participants received US$10 as an incentive for their participation. At the second university, students voluntarily entered a raffle to win one of eight US$10 gift certificates. Participants read and signed the consent form and took a 30-min online survey in English. The study was approved by the university's institutional review board." (Kalibatseva & Leong, 2018, p. 1560)
Sharing/Openness (i.e., materials, data, code).	2	Materials: "*Demographic questionnaire*. Demographic information was collected on age, gender (0 = *male*, 1 = *female*), race (0 = *Chinese American*, 1 = *European American*), class standing, income (rated on a Likert-type scale from 1 to 11, where 1 = US$0 to US$14,999 and 11 = US$105,000 or more), and generational status." (Kalibatseva & Leong, 2018, p. 1560) "*Center for Epidemiological Studies Depression Scale (CES-D)*. The CES-D measures the frequency of 20 symptoms of depression over the past week. It uses a 4-point Likert-type scale ranging from 0 (rarely or none of the time) to 3 (most or all of the time) and higher scores indicate higher levels of depression. The CES-D has four subscales: affective, somatic, positive, and interpersonal

		(Hales et al., 2006).” (Kalibatseva & Leong, 2018, p. 1560) “*Patient Health Questionnaire-15 (PHQ-15).* The PHQ-15 is a self-report questionnaire that measures the severity of 15 somatic symptoms over the past 4 weeks (Kroenke, Spitzer, & Williams, 2002). It is a widely used screening instrument for somatization syndromes.” (Kalibatseva & Leong, 2018, p. 1561) “*The Self-Construal Scale (SCS).* The SCS (Singelis, 1994) assesses independent and interdependent self-construal. It consists of two scales with 12 items, each rated on a 7-point Likert-type scale from 1 (*strongly disagree*) to 7 (*strongly agree*).” (Kalibatseva & Leong, 2018, p. 1561) “Loss of Face (LOF) Scale. Participants completed the LOF scale (Zane, 2000; Zane & Yeh, 2002) that contains 21 items measuring a person’s self-assessment of sensitivity to face loss in different situations.” (Kalibatseva & Leong, 2018, p. 1561) “*Emotion Regulation Questionnaire (ERQ).* This 10-item self-report scale was designed to measure respondents’ tendency to regulate their emotions (Gross & John, 2003). It consists of two subscales that measure CR and ES with both positive and negative tone items.” (Kalibatseva & Leong, 2018, p. 1561-1562)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	“Based on multiple regression analyses” (Kalibatseva & Leong, 2018, p. 1556).
Fallacious interpretation of (lack of) statistical significance.	0	Effect sizes: Cohen’s *d* in Table 1 and partial η^2 in Table 2 (Kalibatseva & Leong, 2018, p. 1563). “Based on multiple regression analyses, European American students reported higher levels of somatic symptoms on the Patient Health Questionnaire–15 (PHQ-15) than Chinese Americans. There was no initial group difference in depressive symptoms based on Center for Epidemiologic Studies–

Depression Scale (CES-D) scores. Correlations between depressive and somatic symptoms, independent and interdependent self-construal, and cognitive reappraisal and independent self-construal were stronger for European Americans than Chinese Americans. Somatic symptoms, loss of face, and expressive suppression were positively associated with depressive symptoms, whereas independent self-construal and cognitive reappraisal were negatively associated with depressive symptoms for both groups. When controlling for gender and somatic symptoms, being Chinese American and male was significantly and positively associated with depressive symptoms measured with the CES-D. These ethnic and gender differences in depressive symptoms were explained by independent self-construal, loss of face, cognitive reappraisal, and expressive suppression" (Kalibatseva & Leong, 2018, p. 1556)

"Hypothesis 1 was tested with multiple regressions controlling for gender, age, class, and income. The results revealed that European Americans reported higher somatic symptom (PHQ-15) scores than Chinese Americans (Table 3) and there was no difference in total CES-D scores or the somatic depressive CES-D subscales (Tables 4 and 5). The subscale CES-D analyses were performed because of possible response style bias on the CES-D positive subscale (e.g., Li & Hicks, 2010). Indeed, there was a significant difference in the CES-D positive subscale with Chinese Americans reporting higher scores than European Americans after the items were reverse-coded.

For Hypothesis 2, Pearson's correlations for each sample and comparisons using Fisher r-to-z transformation and two-tailed significance tests (Meng, Rosenthal, & Rubin, 1992) showed that four correlations significantly differed between the two samples (see Table 6). (…)

To test Hypothesis 3, a hierarchical linear regression was used with demographics (gender, age, class, and income) and ethnicity entered in Step 1, somatic symptoms in Step 2, and self-construal, loss of face, and emotion regulation in Step 3. Table 7 shows the results, indicating that the predictors explained 31.8% of the variance in depressive symptoms. Somatic symptoms, loss of face, and ES were positively associated with depressive symptoms, whereas independent self-construal and cognitive reappraisal were negatively

associated with depression. Ethnicity and gender predicted depressive symptoms in Step 2 but this was no longer the case when the psychological constructs (i.e., self-construal, loss of face, and emotion regulation) were added in Step 3.[2]" (Kalibatseva & Leong, 2018, p. 1562-1564)

"There was no evidence that somatization defined in this way was more prominent among Chinese American college students. These findings go against the proposition that Chinese Americans may somatize distress more than European Americans by reporting somatic symptoms in place of affective depressive symptoms. The present results are consistent with the relatively scarce literature that Chinese Americans are not more likely to report higher levels of somatic complaints than European Americans (Mak & Zane, 2004; Ryder et al., 2008). In fact, in the current study, identifying as European American and female was associated with more somatic complaints. It is important to note that Chinese Americans reported lower levels of positive affect than European Americans, which has previously been discussed as a potential explanation for elevated CES-D scores for this group (Li & Hicks, 2010; Ying, 1988). Yet, in this study, there was no difference in overall depressive symptom scores." (Kalibatseva & Leong, 2018, p. 1565-1566)

"Only loss of face was endorsed more strongly by Chinese American college students compared with European American college students and this difference had a small effect size (d = .17). Although ES had a similar effect size (d = .16) and the Chinese American sample endorsed it at a higher level than the European American sample, this difference did not reach statistical significance (p = .07).
Correlations among the tested variables were largely similar. Significance testing showed there were four correlation coefficients that were significantly different between the two samples. First, the relationship between depressive and somatic symptoms was stronger among European Americans (r = .47) than Chinese Americans (r = .30). (…) Thus, Hypothesis 2 was only partially supported.
A secondary goal of the study was to compare the interconstruct relationships between samples. Whereas most of the correlations were similar for

		the two groups, one difference was that independent and interdependent self-construal correlated more strongly in European Americans (r = .52) than Chinese Americans (r = .20). This finding suggests that there may be a different relationship between the two types of self-construal, such that European Americans may not differentiate between the two or do not find them conflicting in the same way Chinese Americans might. However, for Chinese Americans, independent and interdependent self-construal may be less connected and more differentiated (Markus & Kitayama, 1991). Furthermore, there was a stronger relationship between cognitive reappraisal and independent self-construal among European Americans and a stronger relationship between loss of face and ES among Chinese Americans. These findings may indicate that emotion regulation strategies, such as reinterpreting the meaning of emotion stimuli and suppressing emotions have different associations with how one defines oneself and protects oneself from losing respect and status in one's group." (Kalibatseva & Leong, 2018, p. 1566) "A hierarchical regression analysis revealed that when somatic symptoms were accounted for, a difference in depressive symptoms was evident with Chinese Americans scoring higher than European Americans. This ethnic difference disappeared in the third step after self-construal, loss of face, and emotion regulation were added. Gender and ethnicity were no longer significant predictors of depressive symptoms in Step 3, suggesting that the culturally relevant variables explained existing demographic differences. Gender, ethnicity, and their interaction played an important role on somatic and depressive symptoms among Chinese American and European American college students. In particular, this study found that European American females reported the highest levels of somatic symptoms compared with all other groups. There was also a significant interaction of gender and ethnicity in depressive symptoms. This interaction needs to be interpreted with caution and further research is needed." (Kalibatseva & Leong, 2018, p. 1567)
Assessing the evidential value of a single article by judging the single-	0	"Based on independent *t* tests and chi-square tests, the two samples differed in generation status, age, class standing, and income (see Table 1). Whereas the two samples were comparable in terms of

article *p*-curve (Simonsohn et al., 2014).

gender distribution, there were gender differences in one of the outcome variables. In particular, women (M = 6.19, SD = 4.08) had higher levels of somatic symptoms on the PHQ-15 than men, M = 4.03; SD = 3.92; $t(516) = -5.93$, $p < .001$, consistent with previous research (Kroenke & Spitzer, 1998). There was no ethnic difference in depressive symptoms on the CES-D alone. To disentangle the role of gender and ethnicity on the outcome variables further, a 2 × 2 MANOVA examined the effects of gender and ethnicity on somatic symptoms (PHQ-15) and depressive symptoms (CES-D) together (Table 2). Results revealed significant main effects for gender and ethnicity and a significant interaction (Gender × Ethnicity) for somatic symptoms (PHQ-15) and depressive symptoms (CES-D). In particular, post hoc Tukey tests showed that European American females reported higher somatic symptom (PHQ-15) scores than any of the other three groups ($p < .01$). Chinese American males reported the highest CES-D scores compared with the other three groups. However, post hoc Tukey tests revealed that this difference did not reach statistical significance ($p = .057$). Generational status was not controlled because the study proposed to test the incremental value of ethnicity as a demographic predictor along with culturally relevant predictors in Hypothesis 3.[1]" (Kalibatseva & Leong, 2018, p. 1562)

The first footnote: "When generational status was examined as a predictor of depressive (Center for Epidemiologic Studies–Depression Scale [CES-D]) and somatic (Patient Health Questionnaire-15 [PHQ-15]) symptoms among the Chinese American participants only in a MANOVA, there were no statistically significant differences, $F(4, 372) = 1.18$, $p = .32$; Wilks's $\Lambda = .98$, partial $\eta^2 = .01$." (Kalibatseva & Leong, 2018, p. 1568)

The second footnote: "The hierarchical regression was also performed with CES-D without including the positive affect subscale as it has been problematic with Chinese and Chinese American participants. The results were similar to those reported in Table 7: Step 1: $R^2 = .004$, $F(5, 466) = .38$ (*ns*); Step 2: $\Delta R^2 = .19$, $F(6, 465) = 18.99$**; Step 3: $\Delta R^2 = .09$, $F(11, 460) = 16.68$**." (Kalibatseva & Leong, 2018, p. 1569)

In Figure D9, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $t(516) = -5.93$; $F(4, 372) = 1.18$.

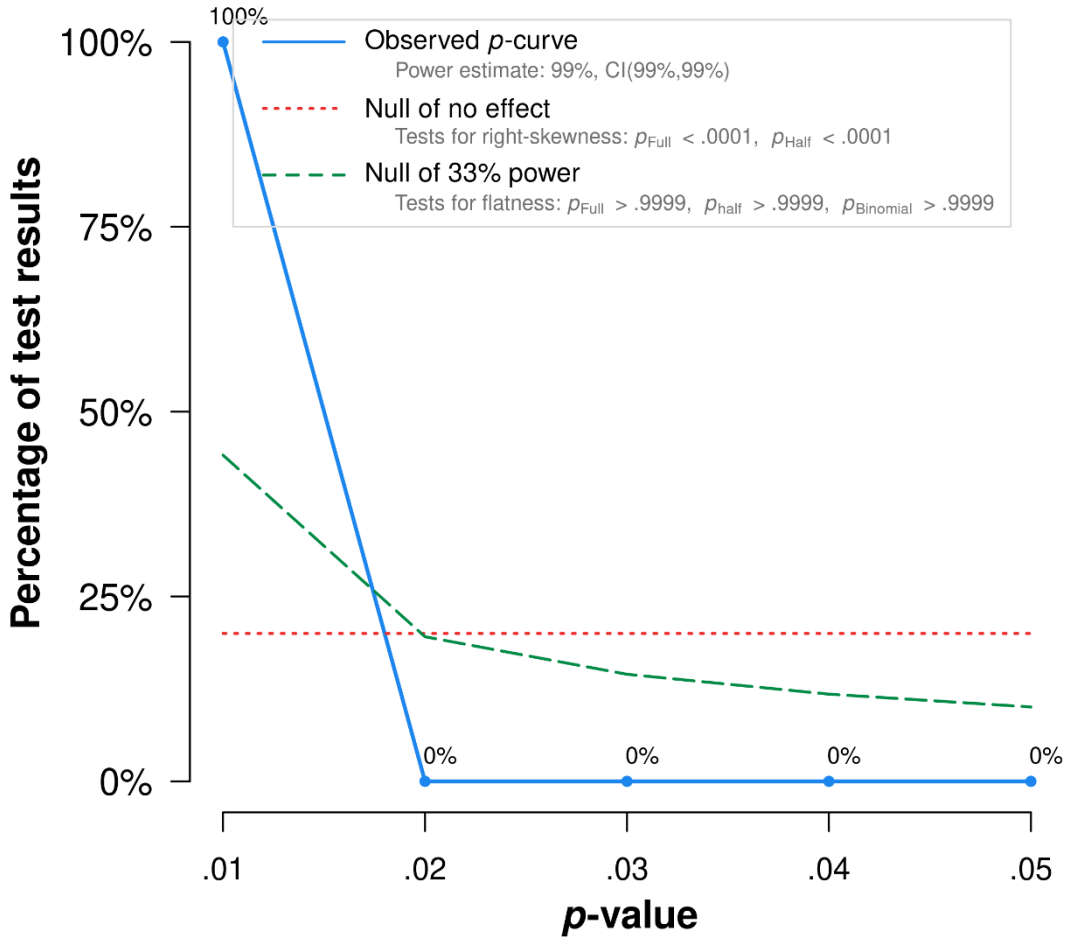


Figure D9. The single-article *p*-curve for the number 7 of the bottom 10 (i.e., Kalibatseva & Leong, 2018).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D10, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p = .0003$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power

test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D10, the 33% power test is $p > .9999$ for the full *p*-curve, for the half *p*-curve, and for the binomial 33% power test; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	p=.5	Z=-5.18, p<.0001	Z=-5.05, p<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	p>.9999	Z=4.07, p>.9999	Z=4.25, p>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure D10. Additional statistics for single-article p-curve for the number 7 of the bottom 10 (i.e., Kalibatseva & Leong, 2018).

Number 8 of the Bottom 10

The number 8 of the bottom 10 is the first study reported in the paper *An emic-etic approach to personality assessment in predicting social adaptation, risky social behaviors, status striving and social affirmation* (Burtăverde et al., 2018). Table D8 shows the score on each of the eight selected RDF for the number 8 paper.

Table D8

Coding paper nr. 8 of the bottom 10 (i.e., Burtăverde et al., 2018)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Exploratory: "The aims of this research were (1) to utilize a recently developed taxonomy of indigenous (Romanian) personality dispositions relying on the lexical approach (Burtaverde & De Raad, in press) for the construction of a brief measure to assess the trait factors summarizing this taxonomy" (Burtăverde et al., 2018, p. 120) "This research tested the utility of an emic-etic approach to personality assessment in predicting three behavioral domains: Social adaptation, Risky social behaviors and Status striving and social affirmation. In this regard, two studies were conducted. The aim of the first study was to develop a personality measure of a psycho-lexically

based six-factorial trait-taxonomic structure identified in the Romanian lexicon." (Burtăverde et al., 2018, p. 113)

"Since so little is known about the utility of an emic-etic approach, we test the predictive power of the approach using a variety of broad behavioral criteria." (Burtăverde et al., 2018, p. 113)

"The study of Burtăverde and De Raad (in press) ultimately led to a six-factor structure with Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Morality (combining sincerity versus malignancy), and Unconventionality (a version of the Intellect factor). The results of this latter study are used for the development of a personality measure to represent the emic part in the present study. A summary of relevant details on the lexical origin of this measure is provided in the Method section.
To test the incremental validity of the emic or indigenous personality factors we selected behavioral criteria which are known to affect one's personal life from three broad behavioral categories, namely: (1) social adaptation, (2) risky social behaviors, and (3) status striving and social affirmation.
Social adaptation. Some of the broad personality factors, especially Extraversion and Agreeableness, are known to refer to relational and social aspects of one's life, (e.g., De Raad, 1995; Tov, Nai, & Lee, 2016; Trapnell & Wiggin, 1990). Criteria included in this adaptation category should be expected to be predicted by such broad personality factors. We chose the following five behavioral indicators: self-esteem, job satisfaction, career satisfaction, life satisfaction, and perceived stress. (…)
Risky social behaviors. Taking into account that some other broad personality factors, such as Conscientiousness and Honesty-Humility, refer to what degree an individual complies with social norms, values, and principles (e.g., Burtăverde, Chraif, Anitei, & Dumitru, 2017; Schmitt, 2004), we aimed to select criteria that offer information about the extent to which an individual interacts with his or her environment in a maladaptive way. We chose the following four indicators: short mating orientation, previous socio-sexual behavior, safety-related risk activities, and traffic fines. (…)
Status striving and social affirmation. We also intended to select criteria that tell about what guides

		a person's behavior – to what extent is he or she internally or externally motivated. (…) We chose the following four indicators: materialism preference, desire for power, time spent on social networks, and posts on social networks. These criteria refer to individuals characterized by the need of social recognition and validation." (Burtăverde et al., 2018, p. 114) "Based on the preceding literature review and empirical evidence, the aims of this research are: (1) to develop an emic personality measure, that assesses the six personality factors that proceeded from the Burtăverde and De Raad (in press) lexical taxonomy of Romanian personality descriptors, and (2) to assess the predictive validity personality measures, and especially to test the extent to which these emic or indigenous personality factors bring incremental validity in the prediction of external behavioral criteria. For etic measures of personality, use is made of standard Big Five and Big Six instruments." (Burtăverde et al., 2018, p. 120) "**Study 1: The development of an emic personality measure** The starting-point for the personality measure in Romanian was in the psycho-lexically based factor structure, consisting of six factors consisting of indigenous or emic versions of Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Morality, and of Unconventionality." (Burtăverde et al., 2018, p. 114)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	2	"Given the ratings, 34 terms were removed that turned out to have very low means and which were predominantly evaluative in meaning." (Burtăverde et al., 2018, p. 115)
Sample size (predetermined or not).	3	"*Step 2: Structuring the Romanian trait domain* The list with 412 dispositional trait adjectives was administered to 515 participants (430 women and 85 men; mean age 31.4, ranging from 18 to 74) together with a measure of the Big Five (the 50-item IPIP (Goldberg, 1999). The latter instrument was added mainly for the purpose of identification

		of the factors proceeding from the structuring procedure. The participants were instructed to provide self-ratings on both item-lists." (Burtăverde et al., 2018, p. 115)
Sharing/Openness (i.e., materials, data, code).	2	Materials are shared in the Appendix: "*Six-factor model markers identified in the Romanian trait list*" (Burtăverde et al., 2018, p. 122) and "*The 120 lexically based Romanian trait words (those with an * form the 72-item measure*)" (Burtăverde et al., 2018, p. 122). "This part of the study consists of two steps, the first step involving the construction of a representative list of the complete Romanian trait vocabulary, and the second step consisting of the structuring of that representative trait list through the use of Principal Components Analyses on the basis of ratings on those trait descriptors." (Burtăverde et al., 2018, p. 114-115) "*Step 1: Taxonomy* Two people independently scanned a comprehensive Romanian dictionary for adjectives that were relevant for personality trait description. The combination of the two selections was an agreed upon list (Cohen's kappa: 0.89) of 1746 terms. These terms were classified by ten judges according to a system described by Angleitner et al. (1990), which comprised five main categories (Dispositions, Temporary conditions, Social aspects, Overt characteristics, Terms of limited utility). The judges assigned (interjudge reliability: 0.87) 412 terms to the Dispositions-category, to be used for the structuring step." (Burtăverde et al., 2018, p. 115) "By their contents, these six emic factors seem to come close to the factors of the Six-factor model (Ashton et al., 2004). As a further aid in the interpretation of the factors, and to have an indication of the extent to which this Romanian structure corresponds to the Six-factor model, we selected markers of the factors of the Six-factor model (in the Appendix), to correlate the marker-scales with the lexically based factors (using factor scores). We were able to select seven markers for each of the factors of the Six-factor model. The results are given in Table 1, together with the alpha coefficients for the marker-scales. As can be seen, the lexically based C, ES, and U factors corresponded quite well with Six-factor model

		Conscientiousness, Emotionality, and Intellect, respectively, but the other factors did not have unique substantial correlations with a Six-factor marker scale. The multiple correlations tell that the Six-factor model markers were well covered by the lexical factors; Of the lexical factors, however, particularly Morality was not well covered by the Six-factor model markers. In order to have an appropriate emic measure available for further use, we had the six lexically based factors each represented by the 20 highest loading trait adjectives (including both positive and negative terms). These 120 trait-adjectives are listed in the Appendix. Besides this 120-item emic set of items, we also formed a briefer version of this measure. Such a measure is useful in the present and in future studies for purposes of economy. This briefer measure consists of the 12 highest loading traits per factor. These trait-adjectives are indicated by an asterisk in the 120-list in the Appendix. Table 2 gives the correlations between the scales based on these 20 items and the scales based on the 12 items with the original factors, indicating a good representation of the original lexical factors by the marker scales." (Burtăverde et al., 2018, p. 115)
Using covariates and reporting the results with and without the covariates.	0	No covariates.
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	3	"The ratings on the remaining set of 378 terms were ipsatized (within-subject standardized), and subsequently Principal Components Analysis was applied and Varimax rotated. Regarding the number of components to extract we used (1) the ''scree test" (Cattell, 1966) and (2) Horn's parallel analysis based on Monte Carlo simulation (Horn, 1965), both suggesting six components. In addition, we used (3) the Bass-Ackwards procedure (Goldberg, 2006) leading to the hierarchical emergence of components, and (4) interpretability of components. The latter two techniques supported the six-components solution to be accepted as the proper structure to summarize the main contents underlying the Romanian trait population. Those components were labelled (1) Conscientiousness (e.g., *organized, perfectionist, precise versus reckless, careless, disorganized*), (2) Extraversion (e.g., *bold, energetic, dynamic, versus silent, taciturn, pessimistic*), (3) Agreeableness (e.g.,

		sentimental, sensitive, generous, versus arrogant, sly, machiavellian), (4) Emotional Stability (e.g., *calm, temperate, non-aggressive, versus nervous, choleric, irritable*), (5) Morality (e.g., *honest, ethical, fair, versus lascivious, provocative, enticing*), and (6) Unconventionality (e.g., *unconventional, rebellious, disobedient, versus conformist, submissive, conventional, ordinary*). The contents of the these six factors formed the concepts used for the construction of the emic personality measure." (Burtăverde et al., 2018, p. 115)
Fallacious interpretation of (lack of) statistical significance.	2	"The Romanian trait taxonomy revealed a six-factor solution with the factors labelled (1) Conscientiousness, (2) Extraversion, (3) Agreeableness, (4) Emotional stability, (5) Morality, and (6) Unconventionality. This six-factor model resembles to a great extent the Big Five model, the main difference being the presence of the fifth factor which is called Morality, which factor also includes semantics of Honesty of the HEXACO model. The six-factor structure indeed corresponds reasonably well with the six-factors of the HEXACO model, with the weakest link observed between lexical Morality and HEXACO-Honesty. The use of the HEXACO measure next to the Big Five may generally lead to certain complications. As had been argued before (Ashton et al, 2004), the use of the HEXACO model may involve a shift of semantics between Agreeableness and Emotional Stability, in comparison to the Big Five structure. An explanation would be that the HEXACO model contains rotated versions of Big Five Agreeableness and Emotional Stability: individual differences referring to patience and impulsivity are included in Agreeableness instead of in the Emotional Stability corresponding Emotionality factor, as was the case with the Big Five. In the case of Romanian emic factors, however, individual differences specific to patience and impulsivity are included in the factor Emotional stability, at this point agreeing with the Big Five model. We do not discuss the lexical structure of the Romanian personality lexicon in further detail because this was not the main aim of this research. We conducted a lexical approach to obtain an indigenous personality structure in order to test its predictive power, not its resemblance with the Big Five or HEXACO models. Regarding the first aim of the research, we developed a 120- item

		measure to assess the six factors of the Romanian personality lexicon. The factor-structure using these 120 items showed a strong overlap with the initial factorial conceptualization, and with adequate internal consistencies for each factor scale. The convergence between the semantically corresponding scales of the Romanian instrument and those of the BFI-Big Five and the HEXACO were acceptable." (Burtăverde et al., 2018, p. 120-121)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	The paper does not disclose enough statistics to calculate the single-article *p*-curve.

Number 9 of the Bottom 10

The number 9 of the bottom 10 is the first study reported in the paper *Brazilian Adolescents' Just World Beliefs and Its Relationships with School Fairness, Student Conduct, and Legal Authorities* (Thomas & Mucherah, 2018). Table D9 shows the score on each of the eight selected RDF for the number 9 paper.

Table D9

Coding paper nr. 9 of the bottom 10 (i.e., Thomas & Mucherah, 2018)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g. hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)	2	Confirmatory: "We hypothesize that (1) adolescents' Personal and General BJW will predict their perceptions of school fairness; (2) perceived school fairness will mediate the relationship between Personal BJW and student conduct; and (3) perceived school fairness will mediate the relationship between General BJW and perception of legal authorities. These hypotheses will be tested through the construction and comparison of two models, a mediated model and a partially mediated model." (Thomas & Mucherah, 2018, p. 47) "This study extends prior research by including the BJW, school perceptions, compliance, and legal authorities into one model and by testing it in an

alternate cultural context. Findings in this study can serve to validate the relationships tested in Western samples and provide a broader picture of how BJW can influence the perceptions of other contexts compliance with rules. To understand the generalizability of findings and theoretical relationships established in previous research, it is important to see if they hold true in other cultural contexts. There is little literature on Latin American adolescents' BJW or legal socialization, and this study could provide a bridge between cultures to generalize existing knowledge or reveal important contrasts.

Past research has documented relations between BJW and individual behaviors (Dalbert, 2009), between perceptions of school fairness and student conduct (Cohn et al., 2012; Donat et al., 2012) and legal authorities (Gouveia-Pereira et al., 2003), and between BJW and legal authorities (Correia & Vala, 2004). Consolidating these variables into one model will help us understand if the relationship between Personal BJW and student conduct is a direct relationship (as implied by prior research), or if its effect is indirect through the perception of a fair school environment. We hypothesize that both Personal and General BJW relate to perceived school fairness however, since Personal BJW is more closely associated with adaptive outcomes and individual behaviors (Dalbert, 2009), only Personal BJW will predict student conduct." (Thomas & Mucherah, 2018, p. 46)

"The current study aims to tie together research that has independently shown General BJW to influence legal authorities and to shape perceptions of school fairness. These relationships are explained in the subsequent sections." (Thomas & Mucherah, 2018, p. 43)

"Personal BJW is thought to develop early and shaped by parental warmth and consistency (Dalbert & Radant, 2004). BJW has also been shown to influence internal attributions and shape the interpretation of events (Chen & Young, 2013). For these reasons, we expect that students bring their Personal BJW assumptions to school and this helps shape how they make sense of their educational environment. The school may then be an additional filtering mechanism which further influences their motivation to comply with rules and assumptions for the broader systems (e.g., legal

authorities). However, the research has been largely based in European samples, and additional testing of this path is needed in alternative cultural settings." (Thomas & Mucherah, 2018, p. 44)

"While research has independently connected Personal BJW with student conduct (Cohn et al., 2012; Sanches, Gouveia-Pereira, & Carugati, 2012; Way, 2011) and Personal BJW with perceptions of school fairness (Dalbert & Sto¨eber, 2005; Donat et al., 2012; Kamble & Dalbert, 2011; Peter & Dalbert, 2010), the possible direct or indirect relationships have not been analyzed in one cohesive model. This study tests how the perception of school fairness may be an important mediator between Personal BJW and student conduct. Students must perceive they are treated fairly in the school context in order to buy-in collectively to the school rules. In this paper, negative student conduct refers to a disposition of disrespect and non-compliance. It is not meant to refer to typically violent acts or extreme transgressions. But instead, the word is meant to capture minor infractions such as disrespectful comments, parallel conversations, abstaining from constructive participation, and disruptive attitudes. This definition aims to encompass normative development and socialization, not to single out delinquent students or extreme cases." (Thomas & Mucherah, 2018, p. 45)

"this study focuses on students' self-perceived student conduct to understand how it is influenced by their perceptions of justice." (Thomas & Mucherah, 2018, p. 45)

"While Personal BJW may be more strongly activated in personal contexts, General BJW is more closely related to perceptions of distant systems and authorities (Correia & Vala, 2004). In this study, legal authorities are defined as those in the judicial or law enforcement systems. Since most adolescents do not have direct experiences with legal authorities, the broad nature of General BJW may be more relevant in understanding how students create assumptions about the fairness of legal authorities. When students do not believe their school rules are fair, they may build assumptions about society based on the injustices they perceive at school (Cohn et al., 2012)." (Thomas & Mucherah, 2018, p. 46)

		"Adolescents' assumptions about fairness in the world and their experiences at school can sustain or undermine their motivation to comply with school rules. Adolescents who do not perceive the school to be fair, may be less motivated to abide by school rules and may generalize this assumption and assume police will also treat them unfairly. The purpose of this study is to test models of how adolescents' justice perceptions relate to perceptions of school fairness, perceptions of legal authorities, and self-perceived conduct in school. Adolescents' justice judgments are interconnected and dynamic and must be studied in a systemic way to create a more comprehensive model of latent constructs. Analyzing these constructs simultaneously and seeking a model that combines these areas will help us understand the role of the school climate in the creation of an internal working model for justice." (Thomas & Mucherah, 2018, p. 42) "Prior research has demonstrated that adolescence is a sensitive period to develop their belief in a just world (BJW), both general and personal. Research has found significant relationships between BJW, perceptions of school fairness, student conduct, and perceptions of legal authorities. However, no research has combined these constructs in one model to get a broader picture of how adolescents construct their worldview of fairness and how this influences their compliance with authorities." (Thomas & Mucherah, 2018, p. 41)
Exclusion of participants (how many, why, etc.). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.	0	No exclusions.
Sample size (predetermined or not).	3	"Three schools in a city in Southern Brazil participated in the study. There were 137 from the public school, 133 from the private school, and 205 from a military school. (…) The administrations of the schools formally granted access to the classrooms to request student participation during

		class time. Students were informed of the voluntary nature of the study and asked to complete an informed consent form with their parents. The teachers were out of class so as to minimize the social desirability bias. Students took approximately 10–15 min to complete the questionnaire. They then completed the instruments anonymously during class time. Four-hundred and seventy-five students between 8th and 12th grade participated in the study. Of these, 218 (46.1%) were male. The majority of the students self-identified as White (70.1%), reflecting the population statistics of the city (IBGE, 2008). Students ranged from 12 to 19 years old with most being 15 (30.6%) or 16 (23.3%).” (Thomas & Mucherah, 2018, p. 48) “This study analyzed 475 Brazilian adolescents across three schools.” (Thomas & Mucherah, 2018, p. 41) “A Brazilian sample represents an understudied population for just world beliefs scholarship.” (Thomas & Mucherah, 2018, p. 42)
Sharing/Openness (i.e., materials, data, code).	3	“utilizing the R software” (Thomas & Mucherah, 2018, p. 51) “The scales used in this study were published in the English language and translated into Portuguese by a native speaker. Two Brazilian psychologists back-translated the items for authenticity, then two Brazilian teachers analyzed and critiqued the instruments to ensure that all items were at students’ reading levels. The explanation of each specific scale will provide additional information about the translation process. A pilot study was conducted with 47 students at a public school in Southern Brazil to ensure the items were comprehensible and reliable to a Brazilian adolescent population. Students from the pilot study reported to understand the items and had no questions. The pilot study revealed adequate reliability ($\alpha > .60$) of all of the scales. The reliability of each scale as measured in the full sample is reported below.” (Thomas & Mucherah, 2018, p. 48) “**Instruments** Five constructs were measured: Personal BJW, General BJW, perception of school fairness, perceptions of legal authorities, and self-perceived student conduct. All items were measured on a six

		point Likert scale. (…) *Belief in a Just World (BJW)* This was measured through Dalbert's (1999) Personal BJW (seven items; e.g., ''Overall, events in my life are just'') and General BJW questionnaire (six items; e.g., ''I think basically the world is a just place''). (…) The study revealed acceptable reliability estimates for both the Personal BJW (α = .76) and General BJW (α = .65) scales. In addition, confirmatory factor analyses were conducted and revealed to be one dimensional. See the results section for additional details. These items are assessed on a six point Likert scale ranging from 1 = completely disagree to 6 = completely agree. *School Fairness* To measure this, five items from the Delaware School Climate survey (Bear, Gaskins, Blank, & Chen, 2011) (e.g., ''The rules in this school are fair'') and two items from the shortened version of the California School Climate and Safety Survey (Furlong et al., 2005) (e.g., ''It pays to follow the rules at my school'') were used. Items were chosen based on their conceptual relationship with fairness in school. *Perceptions of Legal Authorities* Legal authorities are defined in this measure as authorities in the judicial, legislative, and law enforcement systems. This concept was measured through a compilation of items from scales of evaluation of authorities constructed based on (…) *Student Conduct* Students answered seven items about their level of respect for and compliance with school rules and authorities" (Thomas & Mucherah, 2018, p. 49-50)
Using covariates and reporting the results with and without the covariates.	2	"Preliminary Analysis on Age While the primary purpose of this study is not developmental, it is important to address possible differences across adolescence. To ensure that the tested relationships do not differ substantially across the ages in the sample, bivariate and partial correlations were conducted controlling for age. These were compared and were found to be similar, indicating that controlling for age does not substantially differentiate the relationships between constructs. See Table 2 for the results." (Thomas & Mucherah, 2018, p. 51) "This study will test both the direct relationship between General BJW and perceptions of legal

		authorities, and the indirect relationship through perceived school fairness. Prior research has implied a direct relationship may exist (Correia & Vala, 2004); however, school has also been a significant predictor in past relationships (Gouveia-Pereira et al., 2003) and it theoretically could serve as an internal working model for authorities. Testing the direct and indirect relationships will help establish the variance accounted for by each construct and see if General BJW continues to be significant after accounting for the perception of school fairness." (Thomas & Mucherah, 2018, p. 47)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	1	"**SEM Analysis** To answer the research question, structural equation modeling (SEM) analyses were conducted. A mediated model (Fig. 1, Model A) and a partially mediate model (Fig. 1, Model B) were tested utilizing the R software and the weighted least squares (WLS) estimation method. Prior to testing the hypothesized models, individual confirmatory factor analyses (CFA) were conducted on each construct to investigate whether the data fit the factor structures of the theoretically proposed latent variables. The variables were considered to have adequate fit if the Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI) were ≥ .90 and if the standardized root mean square residual (SRMR) and the root mean square error of approximation were ≤ .08 (Kline, 2011). See Table 3 for the model fit statistics for each latent variable. The latent variables fit as predicted and the analysis of the models proceeded as planned. Each model proposed was analyzed through the R software utilizing the *lavaan* package. For the SEM analysis, χ^2 was not considered an adequate test for model fit because the data are not multivariate normal, an inherent assumption of the χ^2 test. For this reason, the CFI, TLI, RMSEA, and SRMR fit statistics were used. The WLS estimation method was used, which is considered robust to non-normal data (Kline, 2011)." (Thomas & Mucherah, 2018, p. 51-52) "The study tests two models, testing both a direct and indirect relationship. This distinction is important to understand the contributors to adolescents' misconduct." (Thomas & Mucherah, 2018, p. 46)

		"Consistent with structural equation modeling, a composite score was not created, but each item was used to build the tested models." (Thomas & Mucherah, 2018, p. 49) "A partially mediated and a mediated model were tested to determine if students' BJW relate directly or indirectly to student conduct and perceptions of legal authorities through school fairness." (Thomas & Mucherah, 2018, p. 41)
Fallacious interpretation of (lack of) statistical significance.	3	"A strong direct relationship between Personal BJW and student conduct indicates that students' underlying assumptions about the justice in their lives drives their behaviors in school. A strong indirect relationship would suggest that Personal BJW influences adolescents' perceptions of the school, and these environmental perceptions drive the conduct violations. A significant indirect effect would suggest that student conduct problems cannot be reduced to individual ''problem'' students, but must be seen within the context of student beliefs about school rules and authorities (Aquino, 1998)." (Thomas & Mucherah, 2018, p. 46-47) "The partially mediated model best fit the data. Personal BJW predicted students' perceptions of the school fairness, which predicted student conduct. General BJW and school fairness predicted adolescents' perceptions of legal authorities. Perceptions of school fairness are influenced by Personal BJW and are predictive of students' conduct and opinions of legal authorities. By analyzing multiple constructs simultaneously, this study provides a picture of how these overlapping conceptualizations of justice interact. Students who do not believe their school is fair are less likely to respect and abide by the rules and are more likely to also expect unfair treatment from law enforcement and judicial officials. This study points to the importance of students' perceptions of justice at school and highlights the far-reaching implications of students who do not perceive or expect justice in their lives." (Thomas & Mucherah, 2018, p. 41) "Consistent with prior research (Dalbert & Sallay, 2004; Dette, Sto¨eber, & Dalbert, 2004; Sallay, 2004), this analysis revealed that General and Personal BJW are correlated but perform different tasks. The analyses provided a picture of how adolescents construct perceptions of justice within their lives, the school, and the broader community,

		and society. In the sections below, each level of influence is analyzed in light of prior research." (Thomas & Mucherah, 2018, p. 54) "General BJW and perceived school fairness were significantly correlated, but General BJW was not a significant predictor of school fairness in the mediated model and was even weaker in the partially mediated model. (…)The stronger the students' Personal BJW, the more likely they will be to perceive their school to be fair." (Thomas & Mucherah, 2018, p. 55) "we tested the hypothesis that Personal BJW would directly predict student conduct. However, the analysis revealed that the effect of Personal BJW was mediated by school fairness. This finding highlights the importance of justice cognitions within the school context." (Thomas & Mucherah, 2018, p. 55) "The current study initially proposed that General BJW would have a mediating effect on evaluations of legal authorities through perception of school fairness. However, this study suggests that both General BJW and school fairness influence perceptions of legal authorities, but perceived school fairness is not a mediator. Instead, both constructs combine and account for 27% of the variance of legal authorities. Of the models tested, the hypothesized link between General BJW and perceptions of legal authorities is unique to the current research study." (Thomas & Mucherah, 2018, p. 57) "This study is not longitudinal, and no causal conclusions can be drawn from this analysis." (Thomas & Mucherah, 2018, p. 57)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"*Mediated Model* Results from the SEM fit statistics for the mediated model revealed to have good fit: $\chi^2(521) = 895.78$, CFI = .931, TLI = .926, RMSEA [.037–.047], SRMR = .066. The model revealed that Personal BJW significantly predicted students' perceptions of the school fairness ($\beta = .611$, $p < .01$). However, contrary to previous research (Peter & Dalbert, 2010), General BJW did not predict perceived school fairness ($\beta = .136$, $p > .05$). As expected, school fairness significantly predicted perceived justice of legal authorities ($\beta = .426$, $p < .01$) and student conduct

(β = .510, p < .01). Personal BJW had a significant indirect effect on student conduct through school fairness (β = .260, p < .01), but General BJW did not have an indirect effect on perceptions of legal authorities through school fairness (β = .069, p > .05). The model accounted for 18% of the variance in student conduct (R^2 = .181), 45% of the variance in perceptions of school fairness (R^2 = .453), and 26% of the variance of legal authorities (R^2 = .260). See Fig. 2. Note that all path coefficients are standardized.

Partially Mediated Model

The second analysis included both direct and mediated relations between Personal BJW and student conduct, and General BJW and perceptions of legal authorities. Just as the previous model, the model converged after 67 iterations and yielded the same fit because both models account for the same amount of variance. Personal BJW significantly predicted perceptions of school fairness (β = .590, p < .01), but General BJW did not (β = .089, p > .05). Personal BJW did not directly predict student conduct (β = .108, p > .05), but did have a significant indirect effect on student conduct through school fairness (β = .196, p < .01). General BJW significantly predicted perceptions of legal authorities (β = .238, p < .05), but did not have a significant indirect effect on legal authorities through perceived school fairness (β = .036, p > .05). The total effect of Personal BJW, and school fairness on behavior was significant (β = .304, p < .01), and the total effect of General BJW, and school fairness, on authorities was significant (β = .273, p < .05). The model accounted for 16% of the variance of student conduct (R^2 = .166), 39% of the variance of school fairness (R^2 = .396), and 27% of legal authorities (R^2 = .274).

Model Comparison

Using the Chi-square difference test, the two models were compared to understand which model best fit the data. Both models revealed good model fit, but the partially mediated model fits significantly better (p < .05) than the mediated model, χ^2_D = 6.679. The partially mediated model fits the data best because it accounted for the direct relationship between General BJW and legal authorities (β = .238), which was unaccounted for in the mediated model." (Thomas & Mucherah, 2018, p. 52-54)

> "Consistent with prior research, Personal BJW was slightly higher (*M* = 3.97, *SD* = .79) than General BJW (3.31; *SD* = .81). Students moderately endorsed the school fairness construct (*M* = 4.18; *SD* = .94) and mildly endorsed perception of legal authorities (*M* = 3.12; *SD* = .82). On average, students rated themselves moderately abiding by school rules and authorities (*M* = 4.72, *SD* = .73)." (Thomas & Mucherah, 2018, p. 51)

In Figure D11, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $\chi^2(521) = 895.78$.

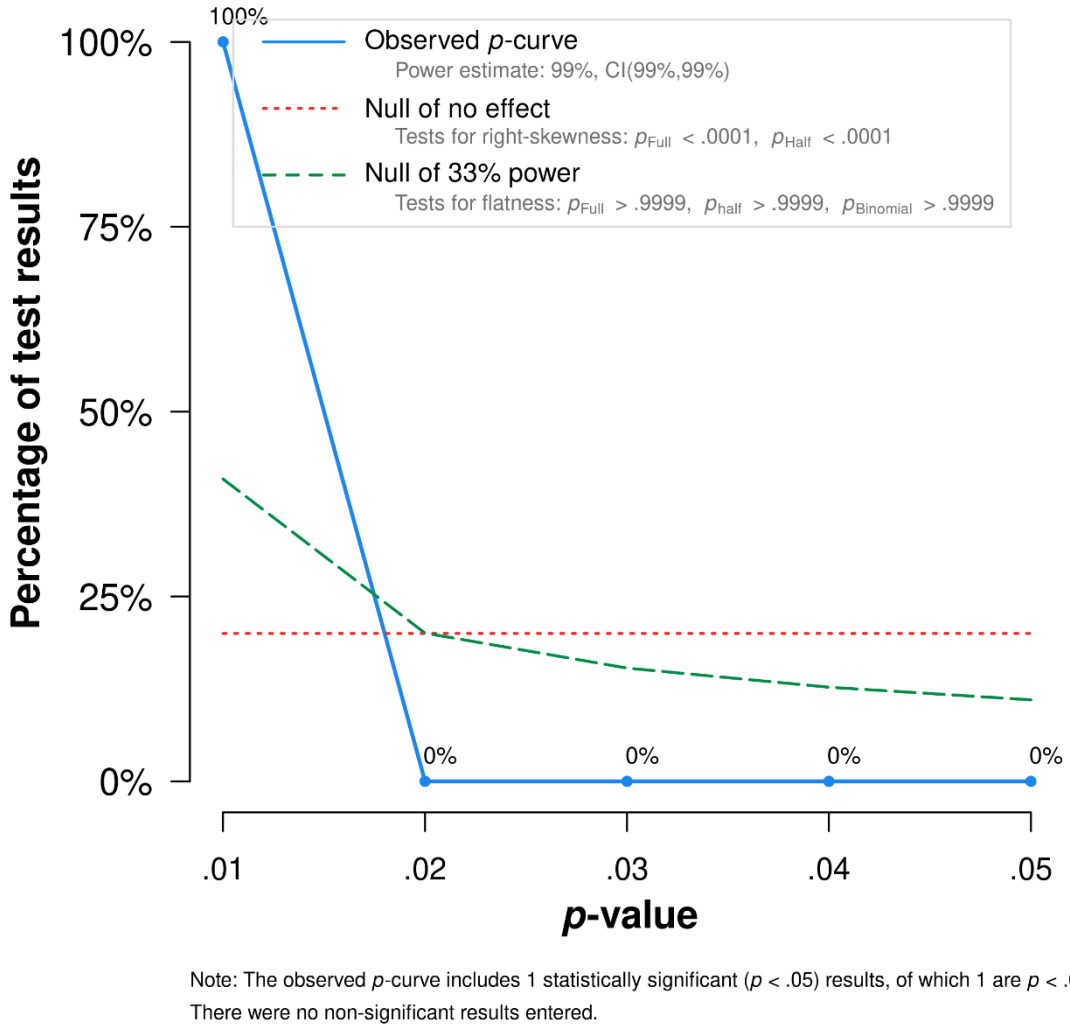


Figure D11. The single-article *p*-curve for the number 9 of the bottom 10 (i.e., Thomas & Mucherah, 2018).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D12, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D12, the 33% power test is $p > .9999$ for the full *p*-curve, for the half *p*-curve, and for the binomial 33% power test; "so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.5	*Z*=-7.76, *p*<.0001	*Z*=-7.67, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*>.9999	*Z*=8.13, *p*>.9999	*Z*=8.13, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure D12. Additional statistics for single-article *p*-curve for the number 9 of the bottom 10 (i.e., Thomas & Mucherah, 2018).

Number 10 of the Bottom 10

The number 10 of the bottom 10 is the first study reported in the paper *Cross-level relationships between justice climate and organizational citizenship behavior: Perceived organizational support as mediator* (Zhang et al., 2017). Table D10 shows the score on each of the eight selected RDF for the number 10 paper.

Table D10

Coding paper nr. 10 of the bottom 10 (i.e., Zhang et al., 2017)

Description RDF	*Score for DFS (0, 1, 2, or 3)*	*Notes*
Confirmatory vs. exploratory (e.g.	2	Confirmatory: "we proposed the following hypotheses:

hypotheses, method, plan of analysis planned beforehand [e.g. preregistration present or clear text indication of divide between planned and unplanned] or ad hoc)		***Hypothesis 1a:*** Procedural justice climate will be positively related to organizational citizenship behavior. ***Hypothesis 1b:*** Interpersonal justice climate will be positively related to organizational citizenship behavior. ***Hypothesis 1c:*** Informational justice climate will be positively related to organizational citizenship behavior." (Zhang et al., 2017, p. 388-389) "In this study, we aggregated the justice judgments of individual employees in the work group, that is, we used the mean individual justice perceptions in the work group to represent the group-level evaluation of fairness. In our model, we assumed that perceived organizational support would play a mediating role in the cross-level relationships between justice climate and organizational citizenship behavior. Thus, we proposed the following hypotheses: ***Hypothesis 2a:*** Perceived organizational support will mediate the cross-level relationships between procedural justice climate and organizational citizenship behavior. ***Hypothesis 2b:*** Perceived organizational support will mediate the cross-level relationships between interpersonal justice climate and organizational citizenship behavior. ***Hypothesis 2c:*** Perceived organizational support will mediate the cross-level relationships between informational justice climate and organizational citizenship behavior." (Zhang et al., 2017, p. 389-390) "We investigated the mediating role of perceived organizational support in the cross-level relationships between procedural, interpersonal, and informational justice climate and organizational citizenship behavior." (Zhang et al., 2017, p. 387) "In this study, we used procedural, interpersonal, and informational justice climate measures to assess the extent to which they may be differentially related to organizational citizenship behavior. (…) Thus, in this study, we sought to identify and examine the mechanisms of the effect of procedural, interpersonal, and informational justice climate on organizational citizenship behavior." (Zhang et al., 2017, p. 388)
Exclusion of participants (how	0	No exclusions.

many, why, etc). Using alternative inclusion and exclusion criteria for selecting participants in analyses. Reporting on how to deal with outliers in an ad hoc manner.		
Sample size (predetermined or not).	3	"We invited staff at 48 hospitals in China to participate in the study. We distributed 600 copies of our survey to the hospitals, of which 468 completed forms were returned, for a 78% response rate. (…) Participants completed the survey during work time in the presence of a researcher. A package containing the survey and a return envelope was given to each participant. In the introductory section of the survey form, the research background was explained and the participants' anonymity was assured. When they had completed the survey, the participants returned their form to the research team in a sealed envelope." (Zhang et al., 2017, p. 390) "Hospital staff in China (N = 468) participated in this study." (Zhang et al., 2017, p. 387)
Sharing/Openness (i.e., materials, data, code).	3	"The measures were translated into Chinese and back to English following Brislin's (1980) back-translation procedure. Ten doctoral students then gave feedback in regard to understanding the questions. As a result, minor adjustments were made to the wording of some of the items. The items were rated on a 5-point Likert-type scale ranging from 1 (*completely disagree*) to 5 (*completely agree*). **Justice climate.** We measured individual perceptions of procedural, interpersonal, and informational justice with items adapted from Colquitt's (2001) scale. (…) **Perceived organizational support.** To measure perceived organizational support, we used four items from the Survey of Perceived Organizational Support developed by Eisenberger et al. (1986). (…) **Organizational citizenship behavior.** We used the six-item scale developed by Wayne, Shore, and Liden (1997) to measure organizational citizenship behavior." (Zhang et al., 2017, p. 390-391)

Using covariates and reporting the results with and without the covariates.	3	"**Control variables.** Because of the potential effect of various demographic variables on the participants' justice perceptions (Caldwell, Liu, Fedor, & Herold, 2009; Lin & Leung, 2014), we controlled for gender, age, education level, and job tenure in our analysis." (Zhang et al., 2017, p. 391)
Reporting completeness on assumption checks. Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner.	2	"Using hierarchical linear modeling" (Zhang et al., 2017, p. 387) "**Data Aggregation** We aggregated participant responses to the organizational level using a referent-shift consensus composition approach recommended by Liao and Rupp (2005). First, we tested the within-group agreement for each type of justice climate by computing the mean r_{wg}. We used a uniform null distribution and found that the mean r_{wg} was .89 for procedural justice climate, .90 for interpersonal justice climate, and .84 for informational justice climate. These results indicated adequate levels of agreement (James, Demaree, & Wolf, 1984). In addition, the intraclass correlation coefficient (ICC) values (1) and (2) for procedural, interpersonal, and informational justice climate were .38 and .86, $F(47, 420) = 7.045$, $p < .001$; .18 and .68, $F(47, 420) = 3.138$, $p < .001$; and .31 and .82, $F(47, 420) = 5.424$, $p < .001$, respectively. Although the ICC (2) value for interpersonal justice climate was lower than ideal, according to the accepted standard, the ICC (1) value was well above the median .12 value used in organizational research, and the F-statistic indicated significant mean differences across groups. Also, the low ICC (2) stemmed in part from the small size of the groups (Bliese, 2000). These results showed that the aggregation of procedural justice climate was justified." (Zhang et al., 2017, p. 391)
Fallacious interpretation of (lack of) statistical significance.	3	"The relationship between procedural ($\gamma = .04$, $p > .05$) and interpersonal ($\gamma = .08$, $p > .05$) justice climate and organizational citizenship behavior, which had previously been significant, then became nonsignificant. These findings suggests that perceived organizational support fully mediated the positive effect of procedural and interpersonal justice climate on organizational citizenship behavior. Therefore, Hypotheses 2a and 2b were supported. As the main effect of informational justice climate on organizational citizenship

		behavior was nonsignificant, Hypothesis 2c was not supported." (Zhang et al., 2017, p. 394) "We found that whereas procedural and interpersonal justice climate predicted organizational citizenship behavior, the informational justice climate relationship with the focused outcomes was nonsignificant. These different effects of the three justice types provide further evidence that the work group perception of the climate of procedural, interpersonal, and informational fairness has a distinct effect on the attitudes and behavior of employees. With respect to the exchange relationship between group members and the organization, not every type of justice can induce employees to behave in ways that enhance organizational effectiveness. Accordingly, other variables may affect the relationship between employees' perception of informational justice climate and organizational citizenship behavior." (Zhang et al., 2017, p. 395) "In this study, we examined the relationship from the perspective of organizational support theory, which is different from the social identity perspective. Overall, our findings suggest that, as a group-level construct, procedural and interpersonal justice climate influence individual behavioral outcomes through individual perceptions of organizational support." (Zhang et al., 2017, p. 395) "although all data were obtained from the same source, the three types of justice climate that we examined were aggregated from individual perceptions, and both the hypotheses involved cross-level relationships. Overall, common method variance was not a substantial concern for our findings (Liao & Rupp, 2005)." (Zhang et al., 2017, p. 396)
Assessing the evidential value of a single article by judging the single-article *p*-curve (Simonsohn et al., 2014).	0	"**Hierarchical Linear Modeling Analysis** **Step 1: Null model.** We used hierarchical linear modeling (HLM 6.08) to test our cross-level hypotheses. First, we estimated a null model to confirm that it was necessary to move to the group level and conduct further cross-level analyses. We found that 15% of the variance resided between groups and the chi-square test result indicated that the between-group variance was significant (χ^2 = 126.36, $p < .001$). These results justified testing our cross-level hypotheses as follows: **Step 2: Random coefficient regression model-1.**

> (…) They showed that both procedural justice climate (γ = .08, p < .05) and interpersonal justice climate (γ = .17, p < .05) had a significantly positive effect on organizational citizenship behavior, supporting Hypotheses 1a and 1b. However, as informational justice climate did not have a significant effect on organizational citizenship behavior (γ = .08, p > .05), Hypothesis 1c was not supported. The results also indicated that procedural (γ = .20, p < .01), interpersonal (γ = .44, p < .001), and informational (γ = .19, p < .01) justice climate all had a significantly positive effect on perceived organizational support. Hence, the first two mediation criteria were met for procedural justice climate and interpersonal justice climate.
> **Step 3: Random coefficient regression model-2.** After we had changed perceived organizational support to the group-mean-centered form in the Level 1 units and included its group mean in Level 2 in Model 2, perceived organizational support was entered into the equation. The relationship between procedural (γ = .04, p > .05) and interpersonal (γ = .08, p > .05) justice climate and organizational citizenship behavior, which had previously been significant, then became nonsignificant." (Zhang et al., 2017, p. 394)

In Figure D13, the results are shown of entering the following statistics into the online *p*-curve app ("*P*-curve app 4.06," 2017): $F(47, 420) = 7.045$; $F(47, 420) = 3.138$; $F(47, 420) = 5.424$.

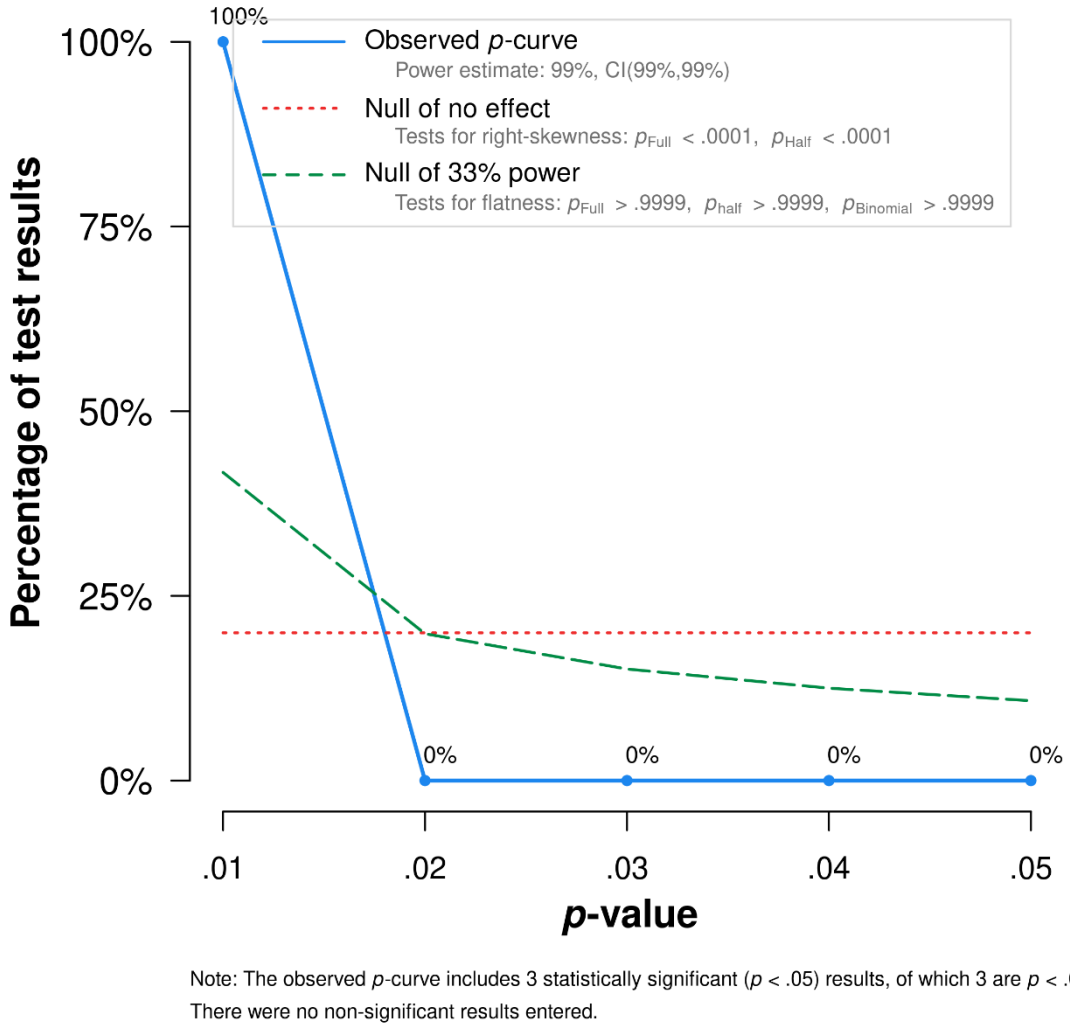


Figure D13. The single-article *p*-curve for the number 10 of the bottom 10 (i.e., Zhang et al., 2017).

According to the combination test of Simonsohn and colleagues (2015, p. 1151), "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half *p*-curves have $p < .1$ right-skew tests." As shown in Figure D14, not only is the half *p*-curve test ($p < .0001$) significantly right-skewed ($p < .05$), but also both the half ($p < .0001$) and full test ($p < .0001$) are significantly right-skewed ($p < .1$), which implies that the study contains evidential value (Simonsohn et al., 2014).

"Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p < .1$." ("*P*-curve results app 4.06," 2017) As shown in Figure D14, the 33% power test is $p > .9999$ for the full *p*-curve, for the half *p*-curve, and for the binomial 33% power test;

"so *p*-curve does not indicate evidential value is inadequate nor absent." ("*P*-curve results app 4.06," 2017)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.125	*Z*=-12.22, *p*<.0001	*Z*=-12.05, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*>.9999	*Z*=9.41, *p*>.9999	*Z*=9.63, *p*>.9999
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99% , 99%)		

Figure D14. Additional statistics for single-article *p*-curve for the number 10 of the bottom 10 (i.e., Zhang et al., 2017).

Appendix E. *R* Code for Reproducing the Current Thesis

This Appendix contains the *R* code for:

- Cleaning the master file;
- Cleaning the extra file;
- Merging the master file with the extra file;
- Completing missing sample sizes after manually looking them up;
- Adding variables for deciding which study has the largest sample within each paper;
- Determining which study within each paper has the largest sample size in said paper;
- Manually fixing the incorrectly coded sample size of study 1 of the paper 'Social Judgment and Social Memory' from 50 to 1185;
- Completing missing citation scores of the extra studies after looking the corresponding papers up in the master file;
- Describing the study numbers, applying the exclusion criteria on the cleaned dataset, and then calculating RV for all studies in the final dataset;
- Generating plots that describe the sample and then for obtaining the top, center, and bottom 10 studies;
- Generating the radar plots for the top, center, and bottom 10 studies.

R Code for Cleaning the Master File

This section contains the *R* code for cleaning the master file. Amongst other cleaning steps:

- Renaming five variables (e.g., more meaningful names, fixing typo's);
- For the DOI variable: Changing the character NA into the missing NA;
- Reassigning 106 variables to the correct class, because a lot of variables are classified as characters while being numeric;
- Nineteen missing DOI's are completed after manually looking them up.

```
# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("naniar") # for using n_miss()
library("writexl")

# Import the raw excel file ----
# path <- insert_your_own_path_to_the_file: "Soc psy Alldata-22-05 copy with sample
size check columns.xlsx"
path <- "C:\\Users\\Celess\\Documents\\Soc psy Alldata-22-05 copy with sample size
check columns.xlsx"

data_raw <- read_excel(path,
                       sheet = 1,
                       guess_max = 21474836)

rm(path)

# Preparing the raw data for further cleaning ----
names(data_raw) # Display all 219 variables of the 999 observations from the raw
data

all.equal(data_raw$...1, data_raw$NA..2) # TRUE, so the first 2 columns are the
same
```

```
all.equal(data_raw$...1, data_raw$NA.) # 247 string mismatches
unique(data_raw$NA.) # Gap between 752 and 760 (so from 752 up until 999: string
mismatches with column 1)

# The following columns can all be changed:
data_1 <- data_raw %>%
  mutate(
    study_number = 1 # Everything from this master file is study 1
  ) %>%
  rename(ID = ...1, # Giving the 1st column a more meaningful name.
         is_longitudinal = is_longtitudinal, # Fixing the typo
         coder_sample_size = `coder_sample size`,  # Underscore instead of space
         prereg = prereg...76,
         prereg_link = prereg_link...77)

data_1$DOI[data_1$DOI=="NA"] <- NA # Changing the character NA into the missing NA

# Overview of the variables and their classes ----
# 220 variables and 999 observations: ----
glimpse(data_1) # A lot of variables are classified as characters while being
numeric

# Therefore, the variables have to be assigned to the correct class: ----
data_2 <- data_1 %>%
  rowwise() %>%
  mutate(
    ID = as.numeric(ID),
    NA. = as.numeric(NA.),
    sample.size.before.exclusion = as.numeric(sample.size.before.exclusion),
    sample_size_after.x = as.numeric(sample_size_after.x),
    coder_sample_size_check = ifelse(coder_sample_size_check %>% tolower %in% 1,
"1", "0"), # NA's are changed into 0
    p_oa = as.numeric(p_oa),
    p_oa_gold = as.numeric(p_oa_gold),
    p_oa_bronze = as.numeric(p_oa_bronze),
    p_oa_hybrid = as.numeric(p_oa_hybrid),
    p_oa_green = as.numeric(p_oa_green),
    mcs = as.numeric(mcs),
    tcs = as.numeric(tcs),
    mncs = as.numeric(mncs),
    mnjs = as.numeric(mnjs),
    pp_top_perc = as.numeric(pp_top_perc),
    pp_uncited = as.numeric(pp_uncited),
    prop_self_cits = as.numeric(prop_self_cits),
    int_cov = as.numeric(int_cov),
    pp_collab = as.numeric(pp_collab),
    pp_int_collab = as.numeric(pp_int_collab),
    NR = as.numeric(NR),
    TC = as.numeric(TC),
    Z9 = as.numeric(Z9),
    U1 = as.numeric(U1),
    U2 = as.numeric(U2),
    SN = as.numeric(SN),
    MF_in_t_a_d_o = as.numeric(MF_in_t_a_d_o),
    MF_clarity = as.numeric(MF_clarity),
    MF_result_report_clarity =as.numeric(MF_result_report_clarity),
    MF_is_positive_finding = as.numeric(MF_is_positive_finding),
    Paper_type = as.numeric(Paper_type),
    is_longitudinal = as.numeric(is_longitudinal),
    paper_clarity = as.numeric(paper_clarity),
    is_replication = as.numeric(is_replication),
    is_exploratory = as.numeric(is_exploratory),
    prereg = as.numeric(prereg),
    prereg_link = as.numeric(prereg_link),
    cohorts.pub = as.numeric(cohorts.pub),
    cohorts.sci = as.numeric(cohorts.sci),
    cohorts.com = as.numeric(cohorts.com),
    context.all.count = as.numeric(context.all.count),
```

```
    context.all.mean = as.numeric(context.all.mean),
    context.all.rank = as.numeric(context.all.rank),
    context.all.pct = as.numeric(context.all.pct),
    context.all.higher_than = as.numeric(context.all.higher_than),
    context.journal.count = as.numeric(context.journal.count),
    context.journal.mean = as.numeric(context.journal.mean),
    context.journal.rank = as.numeric(context.journal.rank),
    context.journal.pct = as.numeric(context.journal.pct),
    context.journal.higher_than = as.numeric(context.journal.higher_than),
    context.similar_age_3m.count = as.numeric(context.similar_age_3m.count),
    context.similar_age_3m.mean = as.numeric(context.similar_age_3m.mean),
    context.similar_age_journal_3m.rank =
as.numeric(context.similar_age_journal_3m.rank),
    context.similar_age_3m.pct = as.numeric(context.similar_age_3m.pct),
    context.similar_age_3m.higher_than =
as.numeric(context.similar_age_3m.higher_than),
    context.similar_age_journal_3m.count =
as.numeric(context.similar_age_journal_3m.count),
    context.similar_age_journal_3m.mean =
as.numeric(context.similar_age_journal_3m.mean),
    context.similar_age_journal_3m.rank =
as.numeric(context.similar_age_journal_3m.rank),
    context.similar_age_journal_3m.pct =
as.numeric(context.similar_age_journal_3m.pct),
    context.similar_age_journal_3m.higher_than =
as.numeric(context.similar_age_journal_3m.higher_than),
    altmetric_id = as.numeric(altmetric_id),
    is_oa = as.logical(is_oa),
    cited_by_posts_count = as.numeric(cited_by_posts_count),
    cited_by_tweeters_count = as.numeric(cited_by_tweeters_count),
    cited_by_accounts_count = as.numeric(cited_by_accounts_count),
    last_updated = as.numeric(last_updated),
    score = as.numeric(score),
    history.1y = as.numeric(history.1y),
    history.6m = as.numeric(history.6m),
    history.3m = as.numeric(history.3m),
    history.1m = as.numeric(history.1m),
    history.1w = as.numeric(history.1w),
    history.6d = as.numeric(history.6d),
    history.5d = as.numeric(history.5d),
    history.4d = as.numeric(history.4d),
    history.3d = as.numeric(history.3d),
    history.2d = as.numeric(history.2d),
    history.1d = as.numeric(history.1d),
    history.at = as.numeric(history.at),
    added_on = as.numeric(added_on),
    published_on = as.numeric(published_on),
    readers.citeulike = as.numeric(readers.citeulike),
    readers.mendeley = as.numeric(readers.mendeley),
    readers.connotea = as.numeric(readers.connotea),
    readers_count = as.numeric(readers_count),
    pmid = as.numeric(pmid),
    cited_by_qna_count = as.numeric(cited_by_qna_count),
    cited_by_fbwalls_count = as.numeric(cited_by_fbwalls_count),
    cited_by_feeds_count = as.numeric(cited_by_feeds_count),
    cited_by_peer_review_sites_count =
as.numeric(cited_by_peer_review_sites_count),
    cited_by_msm_count = as.numeric(cited_by_msm_count),
    cohorts.doc = as.numeric(cohorts.doc),
    cited_by_policies_count = as.numeric(cited_by_policies_count),
    cited_by_gplus_count = as.numeric(cited_by_gplus_count),
    cited_by_wikipedia_count = as.numeric(cited_by_wikipedia_count),
    cited_by_videos_count = as.numeric(cited_by_videos_count),
    cited_by_pinners_count = as.numeric(cited_by_pinners_count),
    cited_by_rdts_count = as.numeric(cited_by_rdts_count),
    Rank_Percentage1 = as.factor(Rank_Percentage1), # 5 levels (i.e., 0-20%; 21-
40%; 41-60%; 61-80%; 81-100%)
```

```
    Rank_Percentage2 = as.factor(Rank_Percentage2), # 5 levels (i.e., 0-20%; 21-
40%; 41-60%; 61-80%; 81-100%)
    Rank_Percentage3a = as.factor(Rank_Percentage3a), # 5 levels (i.e., 0-20%; 21-
40%; 41-60%; 61-80%; 81-100%)
    Rank_Percentage3b = as.factor(Rank_Percentage3b), # 5 levels (i.e., 0-20%; 21-
40%; 41-60%; 61-80%; 81-100%)
    MF_coder = as.factor(MF_coder), # 5 levels (i.e., CK; DE; TvB; TVB; VK)
    In250sample = if_else(In250sample == "yes", 1, 0),
    coder_sample_size = as.factor(coder_sample_size), # 5 levels (i.e., TVB; TvB;
AC; VK; DE)
    standardized.coder.comment = as.factor(standardized.coder.comment) # 8 levels
(i.e., longti; not empirical; not social; other; ptjer; survey; unclear;
unfindable)
  )

data_2 %<>%
  mutate(MF_coder = recode(MF_coder, "TvB" = "TVB"),
         coder_sample_size = recode(coder_sample_size, "TvB" = "TVB")
  )

unique(data_2$In250sample) # NA 1
data_2 %<>%
  mutate(In250sample = replace_na(In250sample, 0)) # 1 if in 250 sample; 0 if not
in 250 sample
unique(data_2$In250sample) # 0 1

# Finding DOIs for cells with missing DOIs in the original data ----
data_2 %<>%
  mutate(DOI = ifelse(Title == "CONTEXTUALIZING SOCIAL SUPPORT: PATHWAYS TO HELP
SEEKING IN LATINOS, ASIAN AMERICANS, AND WHITES",
                      "10.1521/jscp.2014.33.1.1",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "MOOD, MEMORY, AND SOCIAL JUDGMENTS IN CHILDREN",
                      "10.1037/0022-3514.54.4.697",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "SEX-TYPING AND SPATIAL ABILITY - THE ASSOCIATION
BETWEEN MASCULINITY AND SUCCESS ON PIAGET WATER-LEVEL TASK",
                      "10.1007/BF00287356",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "PRODUCTIVE BEHAVIORS OF GLOBAL BUSINESS TEAMS",
                      "10.1016/0147-1767(95)00043-7",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "EFFECTS OF TRANSCENDENTAL MEDITATION VERSUS RESTING
ON PHYSIOLOGICAL AND SUBJECTIVE AROUSAL",
                      "10.1037//0022-3514.44.6.1245",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "PSYCHOPHYSICAL AND SOCIAL RATINGS OF HUMAN-BODY
ODOR",
                      "10.1177/014616727600300126",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "MINORITY INFLUENCE AND PSYCHOSOCIAL IDENTITY",
                      "10.1002/ejsp.2420120405",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "INTERPERSONAL RIGIDITY, HOSTILITY, AND
COMPLEMENTARITY IN MUSICAL BANDS",
                      "10.1037/0022-3514.72.2.362",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "ON NOT BEING ABLE TO AGGRESS",
                      "10.1111/j.2044-8260.1966.tb00966.x",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "EFFECT OF EVALUATION THREAT ON PROCRASTINATION
BEHAVIOR",
                      "10.3200/SOCP.147.3.197-209",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "MODIFICATION OF PRESCHOOL SEX-TYPED BEHAVIORS BY
PARTICIPATION IN ADULT-STRUCTURED ACTIVITIES",
                      "10.1007/BF00287691",
                      ifelse(!is.na(DOI), DOI, NA)),
```

```
         DOI = ifelse(Title == "REM MOTIVATION INDUCED BY BRIEF REM DEPRIVATION -
INFLUENCE OF COGNITION, GENDER, AND PERSONALITY",
                      "10.1037/0022-3514.36.7.741",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "THE EXPERIENCE OF BOREDOM - THE ROLE OF THE SELF-
PERCEPTION OF ATTENTION",
                      "10.1037/0022-3514.57.2.315",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "EFFECTS OF PHYSICAL ATTRACTIVENESS ON HONESTY -
SOCIALLY DESIRABLE RESPONSE",
                      "10.1177/0146167276003001O7",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "SOME OPERATIONALIZATIONS OF THE NEODISSOCIATION
CONCEPT AND THEIR RELATIONSHIP TO HYPNOTIC-SUSCEPTIBILITY",
                      "10.1037//0022-3514.54.6.989",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "SOCIAL CATEGORIZATION, FOCUS OF ATTENTION AND
JUDGMENTS OF GROUP OPINIONS",
                      "10.1111/j.2044-8309.1994.tb01043.x",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "INDIVIDUAL DIFFERENCES IN PERSON MEMORY: THE ROLE
OF SOCIOPOLITICAL IDEOLOGY AND IN-GROUP VERSUS OUT-GROUP MEMBERSHIP IN RESPONSES TO
SOCIALLY RELEVANT BEHAVIOR",
                      "10.1177/0146167296227008",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "BRITISH ETHNOCENTRISM SCALE",
                      "10.1111/j.2044-8260.1967.tb00529.x",
                      ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "CONCEPTUAL SYSTEMS AND HOLLANDS-THEORY OF
VOCATIONAL CHOICE",
                      "10.1037/0022-3514.46.2.376",
                      ifelse(!is.na(DOI), DOI, NA))
  )

# Inspecting the missing data ----
n_miss(data_2)  # 148314 missing values (amongst other things because of empty
variables such as OS_coder)

names(data_2) # 220 variables
identical(data_2$mcs, data_2$tcs) # TRUE, so the variables MCS and TCS are
identical

# Export the clean dataset to Excel ----
# Insert your own path to the file
path <- "C:\\Users\\Celess\\Documents\\Step01_MasterDataClean_V1.xlsx"
write_xlsx(data_2, path)
rm(path)
```

R Code for Cleaning the Extra File

This section contains the *R* code for cleaning the extra file. Amongst other cleaning steps:

- Assigning NA to empty cells;
- Renaming variables to match the names of the master file;
- Manually adding sixteen of the missing DOI's after manually looking them up.

```
# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("naniar") # for using n_miss()
library("writexl")
```

```
# Import the extra data ----
# path <- insert_your_own_path_to_the_file: "met studies 1-2.csv"
path <- "C:\\Users\\Celess\\Documents\\met studies 1-2.csv"

data_raw_extra <- read.csv(path, header = TRUE, sep = ";")
rm(path)

names(data_raw_extra) # 12 variables

# Clean the extra data ----
data_extra1 <- data_raw_extra
data_extra1$AU[data_extra1$AU==""] <- NA # Assigning NA if the cell is empty
data_extra1$TI[data_extra1$TI==""] <- NA # Assigning NA if the cell is empty
data_extra1$PY[data_extra1$PY==""] <- NA # Assigning NA if the cell is empty
data_extra1$DI[data_extra1$DI==""] <- NA # Assigning NA if the cell is empty
data_extra1$UT[data_extra1$UT==""] <- NA # Assigning NA if the cell is empty
data_extra1$exclusion_flagged[data_extra1$exclusion_flagged==""] <- NA # Assigning
NA if the cell is empty
data_extra1$study_number[data_extra1$study_number==""] <- NA # Assigning NA if the
cell is empty
data_extra1$sample_size_before[data_extra1$sample_size_before==""] <- NA #
Assigning NA if the cell is empty
data_extra1$sample_size_after[data_extra1$sample_size_after==""] <- NA # Assigning
NA if the cell is empty
data_extra1$coder[data_extra1$coder==""] <- NA # Assigning NA if the cell is empty
data_extra1$coder_comment[data_extra1$coder_comment==""] <- NA # Assigning NA if
the cell is empty
data_extra1$standardized.coder.comment[data_extra1$standardized.coder.comment==""]
<- NA # Assigning NA if the cell is empty

class(data_extra1$study_number) # character
data_extra1$study_number[data_extra1$study_number=="yes"] <- NA
data_extra1$study_number <- as.numeric(data_extra1$study_number)
class(data_extra1$study_number) # numeric
unique(data_extra1$study_number) # NA 1 2 3 4 5 6 7

data_extra2 <- data_extra1
data_extra2 %<>%
  mutate(Authors = AU,
         AU = NULL,
         Title = TI,
         TI = NULL,
         Publication.Year = PY,
         PY = NULL,
         DOI = DI,
         DI = NULL,
         MF_coder = coder,
         coder = NULL
  )

data_extra2 %<>%
  mutate(sample.size.before.exclusion = sample_size_before,
         sample.size.before.exclusion = as.numeric(sample.size.before.exclusion),
         sample_size_before = NULL,
         sample_size_after.x = sample_size_after,
         sample_size_after.x = as.numeric(sample_size_after),
         sample_size_after = NULL,
         standardized.coder.comment = as.factor(standardized.coder.comment),
         Publication.Year = as.numeric(Publication.Year),
         MF_coder = as.factor(MF_coder)
  )

# Finding DOIs for cells with missing DOIs in the extra data ----
data_extra2 %<>%
  mutate(DOI = ifelse(Title == "CONTEXTUALIZING SOCIAL SUPPORT: PATHWAYS TO HELP
SEEKING IN LATINOS, ASIAN AMERICANS, AND WHITES",
                      "10.1521/jscp.2014.33.1.1",
```

```
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "MOOD, MEMORY, AND SOCIAL JUDGMENTS IN CHILDREN",
                        "10.1037//0022-3514.54.4.697",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "SEX-TYPING AND SPATIAL ABILITY - THE ASSOCIATION
BETWEEN MASCULINITY AND SUCCESS ON PIAGET WATER-LEVEL TASK",
                        "10.1007/BF00287356",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "SELF-SERVING BIASES IN ATTRIBUTION - A BAYESIAN-
ANALYSIS",
                        "10.1037/0022-3514.43.2.197",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "INTERACTIONS AND REINFORCEMENT SENSITIVITY THEORY:
A THEORETICAL ANALYSIS OF RUSTING AND LARSEN (1997)",
                        "10.1016/S0191-8869(98)00019-1",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "PRODUCTIVE BEHAVIORS OF GLOBAL BUSINESS TEAMS",
                        "10.1016/0147-1767(95)00043-7",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "EFFECTS OF TRANSCENDENTAL MEDITATION VERSUS RESTING
ON PHYSIOLOGICAL AND SUBJECTIVE AROUSAL",
                        "10.1037//0022-3514.44.6.1245",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "PSYCHOPHYSICAL AND SOCIAL RATINGS OF HUMAN-BODY
ODOR",
                        "10.1177/014616727600300126",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "MINORITY INFLUENCE AND PSYCHOSOCIAL IDENTITY",
                        "10.1002/ejsp.2420120405",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "IMPAIRED RECALL AND MEMORY DISTURBANCE IN PRESENILE
DEMENTIA",
                        "10.1111/j.2044-8260.1975.tb00151.x",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "INTERPERSONAL RIGIDITY, HOSTILITY, AND
COMPLEMENTARITY IN MUSICAL BANDS",
                        "10.1037/0022-3514.72.2.362",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "ON NOT BEING ABLE TO AGGRESS",
                        "10.1111/j.2044-8260.1966.tb00966.x",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "EFFECT OF EVALUATION THREAT ON PROCRASTINATION
BEHAVIOR",
                        "10.3200/SOCP.147.3.197-209",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "MODIFICATION OF PRESCHOOL SEX-TYPED BEHAVIORS BY
PARTICIPATION IN ADULT-STRUCTURED ACTIVITIES",
                        "10.1007/BF00287691",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "REM MOTIVATION INDUCED BY BRIEF REM DEPRIVATION -
INFLUENCE OF COGNITION, GENDER, AND PERSONALITY",
                        "10.1037/0022-3514.36.7.741",
                        ifelse(!is.na(DOI), DOI, NA)),
         DOI = ifelse(Title == "THE EXPERIENCE OF BOREDOM - THE ROLE OF THE SELF-
PERCEPTION OF ATTENTION",
                        "10.1037/0022-3514.57.2.315",
                        ifelse(!is.na(DOI), DOI, NA))
  )

# Prepare data for merging ----
data_extra3 <- data_extra2

which(duplicated(data_2)) # None
unique(data_2$study_number) # Only study number 1
sum(duplicated(data_2$DOI)) # 19 duplicates on DOI
which(duplicated(data_2$DOI))
sum(is.na(data_2$DOI)) # 20 missing values on DOI
```

```
which(is.na(data_2$DOI)) # All those 19 duplicates on DOI are a duplicate of NA on
DOI. Conclusion: all 999 papers in this dataset are unique, thus study number 1.

which(duplicated(data_extra3)) # Only 1: row 972
unique(data_extra3$study_number) # NA 1 2 3 4 5 6 7
length(which(duplicated(data_extra3$DOI)) ) # 304 duplicates on DOI
sum(is.na(data_extra3$DOI)) # 43 missing values on DOI
which(is.na(data_extra3$DOI))

# Export the cleaned dataset to Excel ----
# Insert your own path to the file
path <- "C:\\Users\\Celess\\Documents\\Step01_ExtraDataClean.xlsx"
write_xlsx(data_extra3, path)
rm(path)
```

R Code for Merging

This section contains the *R* code for merging the master file with the extra file. The new variable *study number* is added to the master file to indicate which study of a paper is meant. Next, the new variable *from extra data* is added to the extra dataset in order to indicate in the merged dataset which observations are originally from the master file (i.e., not from extra data) and which are originally from the extra dataset (i.e., from extra data).

```
# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("naniar") # for using n_miss()
library("writexl")

# Import the clean master file ----
# path <- insert_your_own_path_to_the_file:
"Step01_MasterDataClean_V1.xlsx"
path <- "C:\\Users\\Celess\\Documents\\Step01_MasterDataClean_V1.xlsx"

data_2 <- read_excel(path,
                     sheet = 1,
                     guess_max = 21474836)

rm(path)

# Import the clean extra file ----
# path <- insert_your_own_path_to_the_file:
"Step01_ExtraDataClean.xlsx"
path <- "C:\\Users\\Celess\\Documents\\Step01_ExtraDataClean.xlsx"

data_extra3 <- read_excel(path,
                          sheet = 1,
                          guess_max = 21474836)

rm(path)

# Merge the extra data with the original data ----
# data_2 has 999 observations; data_extra3 has 1252 observations
# Expectation before checking on DOI: merged data has 999 + 1252 = 2251
expected observations (and 220 + 1 (because new variable
```

```
from_extra_data) + 1 (because exclusion_flagged) = 222 expected
variables)
data_extra3 %<>%
  mutate(from_extra_data = 1)

data_merge1 <- full_join(data_2, data_extra3)

which(duplicated(data_merge1)) # Only row 1971, because the other
duplicates have different ID's
unique(data_merge1$study_number) # 1 NA 2 3 4 5 6 7
sum(duplicated(data_merge1$DOI))
View(data_merge1[, c("DOI", "Title", "study_number")])
length(which(duplicated(data_merge1$DOI))) # 1044 duplicates on DOI
sum(is.na(data_merge1$DOI)) # 63 missing values on DOI
which(is.na(data_merge1$DOI))
nrow(distinct(data_merge1, DOI)) # 1207 rows with unique DOI
nrow(distinct(data_merge1, ID)) # 1000 rows with unique ID (because 1 -
999 and NA)

# Assigning ID values to the rows from the extra data ----
data_merge1$from_extra_data[is.na(data_merge1$from_extra_data)] <- 0

data_merge1 %<>% # Based on DOI: papers with the same DOI get the same
ID
  arrange(DOI) %>%
  group_by(DOI) %>%
  mutate(ID = if_else(!is.na(ID), ID,
                      if_else(is.na(DOI), -2, # ID changes from NA to -
2 if DOI is missing
                              if_else(DOI == lag(DOI), lag(ID), -3)
                      )
                  ),
         ID = if_else(!is.na(ID), ID,
                      if_else(is.na(DOI), -2,
                              if_else(DOI == lag(DOI), lag(ID), -3)
                      )
         ),
         ID = if_else(!is.na(ID), ID,
                      if_else(is.na(DOI), -2,
                              if_else(DOI == lag(DOI), lag(ID), -3)
                      )
         ),
         ID = if_else(!is.na(ID), ID,
                      if_else(is.na(DOI), -2,
                              if_else(DOI == lag(DOI), lag(ID), -3)
                      )
         ),
         ID = if_else(!is.na(ID), ID,
                      if_else(is.na(DOI), -2,
                              if_else(DOI == lag(DOI), lag(ID), -3)
                      )
         ),
         ID = if_else(!is.na(ID), ID,
                      if_else(is.na(DOI), -2,
                              if_else(DOI == lag(DOI), lag(ID), -3)
                      )
         ),
         ID = if_else(!is.na(ID), ID,
```

```
                        if_else(is.na(DOI), -2,
                                if_else(DOI == lag(DOI), lag(ID), -3)
                        )
        ),
        ID = if_else(!is.na(ID), ID,
                        if_else(is.na(DOI), -2,
                                if_else(DOI == lag(DOI), lag(ID), -3)
                        )
        )
  )

sum(is.na(data_merge1$ID)) # 233 missing values on ID
length(which(data_merge1$ID == -2)) # ID = -2 appears 43 times
length(which(data_merge1$ID == -3)) # ID = -3 appears 0 times

View(data_merge1[, c("ID", "DOI", "Title", "study_number",
"from_extra_data")])

# Remove papers from extra data that do not have a corresponding study
1 in the original data ----
data_merge2 <- data_merge1 # 2251 obs. of 222 variables

data_merge2 %<>%           # 2018 obs. of 222 variables
  group_by(ID) %>%
  filter(!(is.na(ID) & from_extra_data == 1)) # Remove from dataset if
ID is missing and if origin is extra data

View(data_merge2[, c("ID", "DOI", "Title", "study_number",
"from_extra_data")])
sum(is.na(data_merge2$ID)) # 0 missing values on ID
length(which(data_merge2$ID == -2)) # ID = -2 appears 43 times
length(which(data_merge2$ID == -3)) # ID = -3 appears 0 times
length(which(data_merge2$ID == -4)) # ID = -4 appears 0 times
length(which(data_merge2$ID == -5)) # ID = -5 appears 0 times

data_merge2 %<>%           # 1268 obs. of 222 variables
  group_by(ID) %>%
  filter(!(!is.na(ID) & study_number == 1 & from_extra_data == 1)) #
Remove from dataset if ID is not missing and if study is 1 and if
origin is extra data

data_merge2 %<>%           # 1256 obs. of 222 variables
  group_by(ID) %>%
  filter(!(is.na(DOI) & ID == -2 & is.na(Title) & from_extra_data ==
1)) # Remove if both DOI and Title are missing and if ID = -2 and if
origin is extra data

sum(is.na(data_merge2$ID)) # 0 missing values on "ID"
sum(is.na(data_merge2$DOI)) # 22 missing values on "DOI"
sum(is.na(data_merge2$study_number)) # 0 missing values on "study
number"
sum(is.na(data_merge2$from_extra_data)) # 0 missing values on "from
extra data"
length(which(data_merge2$ID == -2)) # ID = -2 appears 2 times (and both
of those 2 times, the DOI is missing)
length(which(data_merge2$ID == -3)) # ID = -3 appears 0 times
length(which(data_merge2$ID == -4)) # ID = -4 appears 0 times
length(which(data_merge2$ID == -5)) # ID = -5 appears 0 times
```

```
View(data_merge2[, c("ID", "DOI", "Title", "study_number",
"from_extra_data")])

data_merge2 %>%
  filter(is.na(DOI)) %>%
  select("ID", "DOI", "Title", "study_number", "from_extra_data") %>%
  arrange(Title) %>%
  View()

data_merge3 <- data_merge2
data_merge3 %<>% # Based on Title
  arrange(Title) %>%
  group_by(Title) %>%
  mutate(ID = if_else(ID != -2, ID,
                      if_else(is.na(Title), ID,
                              if_else(Title == lag(Title), lag(ID), -5)
                      )
  )
  )

data_merge3 %>%
  filter(is.na(DOI)) %>%
  select("ID", "DOI", "Title", "study_number", "from_extra_data") %>%
  arrange(Title) %>%
  View()

sum(is.na(data_merge3$ID)) # 0 missing values on "ID"
sum(is.na(data_merge3$DOI)) # 22 missing values on "DOI"
sum(is.na(data_merge3$study_number)) # 0 missing values on "study
number"
sum(is.na(data_merge3$from_extra_data)) # 0 missing values on "from
extra data"
length(which(data_merge3$ID == -2)) # ID = -2 appears 0 times
length(which(data_merge3$ID == -3)) # ID = -3 appears 0 times
length(which(data_merge3$ID == -4)) # ID = -4 appears 0 times
length(which(data_merge3$ID == -5)) # ID = -5 appears 0 times

# After merging: still 22 missing values on DOI ----
data_merge4 <- data_merge3
data_merge4 %<>%
  mutate(DOI = ifelse(Title == "A COMPARISON OF ADULT WOMENS AND MENS
ASCRIPTIONS OF NEGATIVE TRAITS TO THE SAME AND OPPOSITE SEX",
                      "10.1007/BF00287365",
                      DOI)
  )

data_merge4 %<>%
  arrange(ID)

# Export the merged dataset to Excel ----
# Insert your own path to the file
path <- "C:\\Users\\Celess\\Documents\\Step01_MergedDataClean_V7.xlsx"
# Merge with clean data
write_xlsx(data_merge4, path)
rm(path)
```

R Code for Completing Sample Sizes

This section contains the *R* code for completing missing sample sizes after manually looking them up.

```
# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("writexl")

# Import the merged dataset from Excel ----
# path <- insert_your_own_path_to_the_file: "Step01_MergedDataClean_V7.xlsx"
path <- "C:\\Users\\Celess\\Documents\\Step01_MergedDataClean_V7.xlsx"

data_1 <- read_excel(path,
                     sheet = 1,
                     guess_max = 21474836)

rm(path)

# Inspecting missing sample sizes ----
length(which(is.na(data_1$sample.size.before.exclusion))) # 268 missing values on
sample.size.before.exclusion
length(which(is.na(data_1$sample_size_after.x))) # 29 missing values on
sample_size_after.x
length(which(is.na(data_1$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.na(data_1$decision_AV_SS_after))) # 1155 missing values on
decision_AV_SS_after
length(which(is.na(data_1$study_number))) # 0 missing values on study_number
length(which(is.na(data_1$from_extra_data))) # 0 missing values on from_extra_data

# Filling in some of the missing sample sizes before exclusion ----
data_1b <- data_1
data_1b %<>%
  mutate(sample.size.before.exclusion =
ifelse(!is.na(sample.size.before.exclusion), sample.size.before.exclusion,
                                       ifelse(is.na(DOI),
sample.size.before.exclusion,
                                              ifelse(DOI !=
"10.1177/014616728062003", # Title: ATTRIBUTING EVIL TO THE SUBJECT, NOT THE
SITUATION - STUDENT REACTION TO MILGRAM FILM ON OBEDIENCE
                                                     sample.size.before.exclusion,
                                                     ifelse(study_number == 2,
106, sample.size.before.exclusion)
                                              )
                                       )
  )
  )

length(which(is.na(data_1b$sample.size.before.exclusion))) # 267 missing values on
sample.size.before.exclusion

# Filling in some of the missing sample sizes after exclusion ----
data_1b %<>%
  mutate(sample_size_after.x = ifelse(!is.na(sample_size_after.x),
sample_size_after.x,
                                    ifelse(is.na(DOI), sample_size_after.x,
                                           ifelse(DOI != "10.1016/0191-
8869(91)90227-3", # Title: EXTENDING REDUCER AUGMENTOR THEORY INTO THE EMOTION
DOMAIN - THE ROLE OF AFFECT IN REGULATING STIMULATION LEVEL
                                                  sample_size_after.x,
```

```
                                                    ifelse(study_number == 2,
sample.size.before.exclusion, sample_size_after.x)
                                            )
                                    )
  ),
  sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                               ifelse(is.na(DOI), sample_size_after.x,
                                      ifelse(DOI != "10.2224/SBP.2015.43.10.1725",
# Title: EFFECTS OF CONGRUENCE OF PRODUCT, VISUAL IMAGE, AND CONSUMER SELF-IMAGE ON
ART INFUSION ADVERTISING
                                             sample_size_after.x,
                                             ifelse(study_number == 2,
sample.size.before.exclusion, sample_size_after.x)
                                      )
                               )
  )
  )

data_1b %<>%
  mutate(
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1177/0146167280062003", #
Title: ATTRIBUTING EVIL TO THE SUBJECT, NOT THE SITUATION - STUDENT REACTION TO
MILGRAM FILM ON OBEDIENCE
                                               sample_size_after.x,
                                               ifelse(study_number == 2,
sample.size.before.exclusion, sample_size_after.x)
                                        )
                                 )
    ),
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1016/S0191-8869(98)00103-
2", # Title: ASSESSMENT OF BELIEFS IN EXERCISE DEPENDENCE: THE DEVELOPMENT AND
PRELIMINARY VALIDATION OF THE EXERCISE BELIEFS QUESTIONNAIRE
                                               sample_size_after.x,
                                               ifelse(study_number == 3, 120,
sample_size_after.x)
                                        )
                                 )
    ),
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1111/J.1559-
1816.1994.TB00558.X", # Title: JUDGED PERSON DANGEROUSNESS AS WEIGHTED AVERAGING
                                               sample_size_after.x,
                                               ifelse(study_number == 2, 25,
sample_size_after.x)
                                        )
                                 )
    ),
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1007/978-3-319-51721-
6_3", # Title: FROM RISK AND TIME PREFERENCES TO CULTURAL MODELS OF CAUSALITY: ON
THE CHALLENGES AND POSSIBILITIES OF FIELD EXPERIMENTS, WITH EXAMPLES FROM RURAL
SOUTHWESTERN MADAGASCAR
                                               sample_size_after.x,
                                               ifelse(study_number == 1, 25,
sample_size_after.x)
                                        )
                                 )
    )
  )

data_1b %<>%
  mutate(sample_size_after.x = ifelse(is.na(DOI), sample_size_after.x,
```

```
                                        ifelse(DOI != "10.1037/H0032863", # Title:
GENERATIONAL CONTINUITY AND DISCONTINUITY IN UNDERSTANDING OF SOCIETAL REJECTION
                                                sample_size_after.x,
                                                ifelse(study_number == 2, 95,
sample_size_after.x)
                                            )
                                        )
  )

data_2 <- data_1b
data_2$sample_size_after.x <- round(data_2$sample_size_after.x, digits = 0)

# Adding new column for final sample size before exclusion based on
decision_AV_SS_before (or else based on sample.size.before.exclusion) ----
data_1 <- data_2

data_1 %<>%
  mutate(final_sample_before = NA,
         final_sample_before = as.numeric(final_sample_before) # Create empty
column
  )

data_1[data_1 == "NA" ] <- NA

data_1 <- data_1[order(data_1$DOI),] # Order data (ascending) on DOI

data_1 %<>%
  group_by(DOI) %>%
  mutate(final_sample_before = if_else(is.na(DOI), -2,
                                       if_else(!is.na(decision_AV_SS_before),
decision_AV_SS_before,

if_else(!is.na(sample.size.before.exclusion), sample.size.before.exclusion,
                                                            -1)
                                       )
  )
  )

View(data_1[, c("Title", "DOI", "decision_AV_SS_before",
"sample.size.before.exclusion", "final_sample_before")])

data_2 <- data_1
data_2 %<>%
  group_by(Title) %>%
  mutate(final_sample_before = if_else(!is.na(DOI), final_sample_before,
                                       if_else(!is.na(decision_AV_SS_before),
decision_AV_SS_before,

if_else(!is.na(sample.size.before.exclusion), sample.size.before.exclusion,
                                                            -1)
                                       )
  )
  )

View(data_2[, c("Title", "DOI", "decision_AV_SS_before",
"sample.size.before.exclusion", "final_sample_before")])
summary(data_2$final_sample_before) # No -2 values on final_sample_before

length(which(is.na(data_2$final_sample_before))) # 0 missing values on
final_sample_before, because they are -1 instead of NA
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.infinite(data_2$final_sample_before))) # 0 inf values on
final_sample_before

data_2 %<>%
  mutate(final_sample_before = na_if(final_sample_before, "-1"))
```

```
length(which(is.na(data_2$final_sample_before))) # 240 missing values on
final_sample_before
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.infinite(data_2$final_sample_before))) # 0 inf values on
final_sample_before

# Adding new column for final sample size after exclusion based on
decision_AV_SS_after (or else based on sample_size_after.x) ----
data_1 <- data_2
data_1 %<>%
  mutate(final_sample_after = NA,
         final_sample_after = as.numeric(final_sample_after), # Create empty column
         decision_AV_SS_after = as.numeric(decision_AV_SS_after) # 3 papers have an
error (because they have comments and not numbers as values) -> they now become NA,
but that is okay because sample size after exclusion is filled in for these 3
papers
  )

data_1 %<>%
  mutate(decision_AV_SS_after = round(decision_AV_SS_after),
         sample_size_after.x = round(sample_size_after.x)
         )

View(data_1[, c("Title", "DOI", "decision_AV_SS_after", "sample_size_after.x",
"final_sample_after")])

data_1[data_1 == "NA" ] <- NA

data_1 <- data_1[order(data_1$DOI),] # Order data (ascending) on DOI

data_1 %<>%
  group_by(DOI) %>%
  mutate(final_sample_after = if_else(is.na(DOI), -2,
                                      if_else(!is.na(decision_AV_SS_after),
decision_AV_SS_after,
                                              if_else(!is.na(sample_size_after.x),
sample_size_after.x,
                                                      -1)
                                      )
  )
  )

View(data_1[, c("Title", "DOI", "decision_AV_SS_after", "sample_size_after.x",
"final_sample_after")])

data_2 <- data_1
data_2 %<>%
  group_by(Title) %>%
  mutate(final_sample_after = if_else(!is.na(DOI), final_sample_after,
                                      if_else(!is.na(decision_AV_SS_after),
decision_AV_SS_after,
                                              if_else(!is.na(sample_size_after.x),
sample_size_after.x,
                                                      -1)
                                      )
  )
  )

View(data_2[, c("Title", "DOI", "decision_AV_SS_after", "sample_size_after.x",
"final_sample_after")])
summary(data_2$final_sample_after) # No -2 values on final_sample_after

length(which(is.na(data_2$final_sample_after))) # 0 missing values on
final_sample_after, because they are -1 instead of NA
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
```

```
length(which(is.na(data_2$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.infinite(data_2$final_sample_after))) # 0 inf values on
final_sample_after

data_2 %<>%
  mutate(final_sample_after = na_if(final_sample_after, "-1"))

length(which(is.na(data_2$final_sample_after))) # 3 missing values on
final_sample_after
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.infinite(data_2$final_sample_after))) # 0 inf values on
final_sample_after

View(data_2[, c("ID", "Title", "DOI", "Authors",
                "sample.size.before.exclusion", "sample_size_after.x",
"decision_AV_SS_before", "decision_AV_SS_after",
                "study_number", "from_extra_data", "exclusion_flagged",
                "MF_coder", "coder_sample_size", "coder_comment",
"standardized.coder.comment", "In250sample",
                "final_sample_before", "final_sample_after")])

# Export the dataset to Excel ----
# path <- insert_your_own_path
path <- "C:\\Users\\Celess\\Documents\\Step01c_FinalSampleSizes_V2.xlsx"
write_xlsx(data_2, path)
rm(path)


# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("magrittr") # for using %<>%
library("writexl")

# Import the dataset from Excel ----
# path <- insert_your_own_path_to_the_file: "data_to_find_missing_ss.csv"
path <- "C:\\Users\\Celess\\Documenten\\data_to_find_missing_ss.csv"
data_1b <- read.csv(path)
rm(path)

# Filling in some of the missing sample sizes ----
data_1b %<>%
  mutate(sample_size_after.x = ifelse(!is.na(sample_size_after.x),
sample_size_after.x,
                                      ifelse(is.na(DOI), sample_size_after.x,
                                             ifelse(DOI != "10.1016/0191-
8869(91)90227-3", # Title: EXTENDING REDUCER AUGMENTOR THEORY INTO THE EMOTION
DOMAIN - THE ROLE OF AFFECT IN REGULATING STIMULATION LEVEL
                                                    sample_size_after.x,
                                                    ifelse(study_number == 2,
sample.size.before.exclusion, sample_size_after.x)
                                                    )
                       )
                       ),
         sample_size_after.x = ifelse(!is.na(sample_size_after.x),
sample_size_after.x,
                                      ifelse(is.na(DOI), sample_size_after.x,
                                             ifelse(DOI !=
"10.2224/SBP.2015.43.10.1725", # Title: EFFECTS OF CONGRUENCE OF PRODUCT, VISUAL
IMAGE, AND CONSUMER SELF-IMAGE ON ART INFUSION ADVERTISING
                                                    sample_size_after.x,
                                                    ifelse(study_number == 2,
sample.size.before.exclusion, sample_size_after.x)
```

```
                                          )
                                    )
        ),
        sample.size.before.exclusion =
ifelse(!is.na(sample.size.before.exclusion), sample.size.before.exclusion,
                                    ifelse(is.na(DOI),
sample.size.before.exclusion,
                                           ifelse(DOI !=
"10.1177/014616728062003", # Title: ATTRIBUTING EVIL TO THE SUBJECT, NOT THE
SITUATION - STUDENT REACTION TO MILGRAM FILM ON OBEDIENCE
                                                  sample.size.before.exclusion,
                                                  ifelse(study_number == 2, 106,
sample.size.before.exclusion)
                                           )
                                    )
        )
  )

data_1b %<>%
  mutate(
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                       ifelse(is.na(DOI), sample_size_after.x,
                                              ifelse(DOI !=
"10.1177/014616728062003", # Title: ATTRIBUTING EVIL TO THE SUBJECT, NOT THE
SITUATION - STUDENT REACTION TO MILGRAM FILM ON OBEDIENCE
                                                     sample_size_after.x,
                                                     ifelse(study_number == 2,
sample.size.before.exclusion, sample_size_after.x)
                                              )
                                       )
    ),
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1016/S0191-8869(98)00103-
2", # Title: ASSESSMENT OF BELIEFS IN EXERCISE DEPENDENCE: THE DEVELOPMENT AND
PRELIMINARY VALIDATION OF THE EXERCISE BELIEFS QUESTIONNAIRE
                                               sample_size_after.x,
                                               ifelse(study_number == 3, 120,
sample_size_after.x)
                                        )
                                 )
    ),
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1111/J.1559-
1816.1994.TB00558.X", # Title: JUDGED PERSON DANGEROUSNESS AS WEIGHTED AVERAGING
                                               sample_size_after.x,
                                               ifelse(study_number == 2, 25,
sample_size_after.x)
                                        )
                                 )
    ),
    sample_size_after.x = ifelse(!is.na(sample_size_after.x), sample_size_after.x,
                                 ifelse(is.na(DOI), sample_size_after.x,
                                        ifelse(DOI != "10.1007/978-3-319-51721-
6_3", # Title: FROM RISK AND TIME PREFERENCES TO CULTURAL MODELS OF CAUSALITY: ON
THE CHALLENGES AND POSSIBILITIES OF FIELD EXPERIMENTS, WITH EXAMPLES FROM RURAL
SOUTHWESTERN MADAGASCAR
                                               sample_size_after.x,
                                               ifelse(study_number == 1, 25,
sample_size_after.x)
                                        )
                                 )
    )
  )

data_1b %<>%
  mutate(ss = ifelse(!is.na(ss), ss,
```

```
                    ifelse(DOI != "10.1016/0191-8869(91)90227-3", # Title:
EXTENDING REDUCER AUGMENTOR THEORY INTO THE EMOTION DOMAIN - THE ROLE OF AFFECT IN
REGULATING STIMULATION LEVEL
                    ss,
                    ifelse(study_number == 2, sample_size_after.x, ss))
                    ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(DOI != "10.2224/SBP.2015.43.10.1725", # Title: EFFECTS
OF CONGRUENCE OF PRODUCT, VISUAL IMAGE, AND CONSUMER SELF-IMAGE ON ART INFUSION
ADVERTISING
                           ss,
                           ifelse(study_number == 2, sample_size_after.x, ss))
        ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(DOI != "10.1177/014616728062003", # Title: ATTRIBUTING
EVIL TO THE SUBJECT, NOT THE SITUATION - STUDENT REACTION TO MILGRAM FILM ON
OBEDIENCE
                           ss,
                           ifelse(study_number == 2, sample_size_after.x, ss))
        ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(DOI != "10.1016/S0191-8869(98)00103-2", # Title:
ASSESSMENT OF BELIEFS IN EXERCISE DEPENDENCE: THE DEVELOPMENT AND PRELIMINARY
VALIDATION OF THE EXERCISE BELIEFS QUESTIONNAIRE
                           ss,
                           ifelse(study_number == 3, sample_size_after.x, ss))
        ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(DOI != "10.1111/J.1559-1816.1994.TB00558.X", # Title:
JUDGED PERSON DANGEROUSNESS AS WEIGHTED AVERAGING
                           ss,
                           ifelse(study_number == 2, sample_size_after.x, ss))
        ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(DOI != "10.1007/978-3-319-51721-6_3", # Title: FROM
RISK AND TIME PREFERENCES TO CULTURAL MODELS OF CAUSALITY: ON THE CHALLENGES AND
POSSIBILITIES OF FIELD EXPERIMENTS, WITH EXAMPLES FROM RURAL SOUTHWESTERN
MADAGASCAR
                           ss,
                           ifelse(study_number == 1, sample_size_after.x, ss))
        ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(Title != "PLATE TOUCHING IN RESTAURANTS - PRELIMINARY-
OBSERVATIONS OF A FOOD-RELATED MARKING BEHAVIOR IN HUMANS",
                           ss,
                           ifelse(study_number == 1, sample_size_after.x, ss))
        ),
        ss = ifelse(!is.na(ss), ss,
                    ifelse(Title != "CHANGING SELF-ACCEPTANCE - TASK GROUPS AND
VIDEO TAPE FEEDBACK OR SENSITIVITY TRAINING",
                           ss,
                           ifelse(study_number == 1, sample_size_after.x, ss))
        )
  )

data_2 <- data_1b
data_2$sample_size_after.x <- round(data_2$sample_size_after.x, digits = 0)
data_2$ss <- round(data_2$ss, digits = 0)

# Export the dataset to Excel ----
# path <- insert_your_own_path
path <- "C:\\Users\\Celess\\Documents\\Step01b_MissingSampleSizes_V1.xlsx"
write_xlsx(data_2, path)
rm(path)
```

R Code for Adding the Largest Samples

This section contains the *R* code for adding the following variables for deciding which study has the largest sample within each paper:

- Final sample size before exclusion. Based on the variable "decision_AV_ss_before". If "decision_AV_ss_before" is unknown, then based on the variable "sample.size.before.exclusion".
- Final sample size after exclusion. Based on the variable "decision_AV_ss_after". If "decision_AV_ss_after" is unknown, then based on the variable "sample.size.after.exclusion".
- Largest sample before. Using the variable "final_sample_before" to decide per paper how big the sample size of the study with the largest sample before exclusion(s) within that paper is.
- Largest sample after. Using the variable "final_sample_after" to decide per paper how big the sample size of the study with the largest sample after exclusion(s) within that paper is.

For example, if a paper reports about one single study, then the sample size of that study is automatically the largest sample size of said paper.

```
# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("writexl")

# Import the merged dataset from Excel ----
# path <- insert_your_own_path_to_the_file: "Step01c_FinalSampleSizes_V2.xlsx"
path <- "C:\\Users\\Celess\\Documents\\Step01c_FinalSampleSizes_V2.xlsx"

data_1 <- read_excel(path,
                     sheet = 1,
                     guess_max = 21474836)

rm(path)

# Adding new column for final sample size before exclusion based on
decision_AV_SS_before (or else based on sample.size.before.exclusion) ----
data_1 %<>%
  mutate(final_sample_before = NA,
         final_sample_before = as.numeric(final_sample_before) # Create empty
column
  )

data_1[data_1 == "NA" ] <- NA

data_1 <- data_1[order(data_1$DOI),] # Order data (ascending) on DOI

data_1 %<>%
  group_by(DOI) %>%
  mutate(final_sample_before = if_else(is.na(DOI), -2,
                                       if_else(!is.na(decision_AV_SS_before),
decision_AV_SS_before,

if_else(!is.na(sample.size.before.exclusion), sample.size.before.exclusion,
                                                            -1)
                                       )
```

```
  )
  )

View(data_1[, c("Title", "DOI", "decision_AV_SS_before",
"sample.size.before.exclusion", "final_sample_before")])

data_2 <- data_1
data_2 %<>%
  group_by(Title) %>%
  mutate(final_sample_before = if_else(!is.na(DOI), final_sample_before,
                                       if_else(!is.na(decision_AV_SS_before),
decision_AV_SS_before,

if_else(!is.na(sample.size.before.exclusion), sample.size.before.exclusion,
                                                             -1)
                                       )
  )
  )

View(data_2[, c("Title", "DOI", "decision_AV_SS_before",
"sample.size.before.exclusion", "final_sample_before")])
summary(data_2$final_sample_before) # No -2 values on final_sample_before

length(which(is.na(data_2$final_sample_before))) # 0 missing values on
final_sample_before, because they are -1 instead of NA
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.infinite(data_2$final_sample_before))) # 0 inf values on
final_sample_before

data_2 %<>%
  mutate(final_sample_before = na_if(final_sample_before, "-1"))

length(which(is.na(data_2$final_sample_before))) # 240 missing values on
final_sample_before
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.infinite(data_2$final_sample_before))) # 0 inf values on
final_sample_before

# Adding new column for final sample size after exclusion based on
decision_AV_SS_after (or else based on sample_size_after.x) ----
data_1 <- data_2
data_1 %<>%
  mutate(final_sample_after = NA,
         final_sample_after = as.numeric(final_sample_after), # Create empty column
         decision_AV_SS_after = as.numeric(decision_AV_SS_after) # 3 papers have an
error, because they have comments and not numbers as values
  )

data_1 %<>%
  mutate(decision_AV_SS_after = round(decision_AV_SS_after),
         sample_size_after.x = round(sample_size_after.x)
         )

View(data_1[, c("Title", "DOI", "decision_AV_SS_after", "sample_size_after.x",
"final_sample_after")])

data_1[data_1 == "NA" ] <- NA

data_1 <- data_1[order(data_1$DOI),] # Order data (ascending) on DOI

data_1 %<>%
  group_by(DOI) %>%
  mutate(final_sample_after = if_else(is.na(DOI), -2,
```

```
                                        if_else(!is.na(decision_AV_SS_after),
decision_AV_SS_after,
                                                if_else(!is.na(sample_size_after.x),
sample_size_after.x,
                                                        -1)
                                        )
  )
  )

View(data_1[, c("Title", "DOI", "decision_AV_SS_after", "sample_size_after.x",
"final_sample_after")])

data_2 <- data_1
data_2 %<>%
  group_by(Title) %>%
  mutate(final_sample_after = if_else(!is.na(DOI), final_sample_after,
                                        if_else(!is.na(decision_AV_SS_after),
decision_AV_SS_after,
                                                if_else(!is.na(sample_size_after.x),
sample_size_after.x,
                                                        -1)
                                        )
  )
  )

View(data_2[, c("Title", "DOI", "decision_AV_SS_after", "sample_size_after.x",
"final_sample_after")])
summary(data_2$final_sample_after) # No -2 values on final_sample_after

length(which(is.na(data_2$final_sample_after))) # 0 missing values on
final_sample_after, because they are -1 instead of NA
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.infinite(data_2$final_sample_after))) # 0 inf values on
final_sample_after

data_2 %<>%
  mutate(final_sample_after = na_if(final_sample_after, "-1"))

length(which(is.na(data_2$final_sample_after))) # 3 missing values on
final_sample_after
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.infinite(data_2$final_sample_after))) # 0 inf values on
final_sample_after


# Adding new columns for indicating per paper what is the largest sample size
(before and after exclusion) of their study/studies ----
data_1 <- data_2
data_1 %<>%
  mutate(largest_sample_before = NA,
         largest_sample_before = as.numeric(largest_sample_before), # Create empty
column
         largest_sample_after = NA,
         largest_sample_after = as.numeric(largest_sample_after) # Create empty
column
  )

data_1[data_1 == "NA" ] <- NA

data_1 <- data_1[order(data_1$DOI),] # Order data (ascending) on DOI

## Using final_sample_before to determine per paper what is the largest sample size
before  ----
data_2 <- data_1
```

```
data_2 %<>%
  group_by(DOI) %>%
  mutate(largest_sample_before = if_else(is.na(DOI), -2,
                                         if_else(!is.na(final_sample_before),
as.numeric(max(final_sample_before,na.rm = TRUE)),
                                                 -1
                                         )
  )
  )

data_2 %<>%
  group_by(Title) %>%
  mutate(largest_sample_before = if_else(!is.na(DOI), largest_sample_before,
                                         if_else(!is.na(final_sample_before),
as.numeric(max(final_sample_before,na.rm = TRUE)),
                                                 -1
                                         )
                                         )
  )

summary(data_2$largest_sample_before) # No -2 values on largest_sample_before

length(which(is.na(data_2$largest_sample_before))) # 0 missing values on
largest_sample_before, because they are -1 instead of NA
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.infinite(data_2$largest_sample_before))) # 0 inf values on
largest_sample_before

data_2 %<>%
  mutate(largest_sample_before = na_if(largest_sample_before, "-1"))

length(which(is.na(data_2$largest_sample_before))) # 240 missing values on
largest_sample_before
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.infinite(data_2$largest_sample_before))) # 0 inf values on
largest_sample_before

## Using final_sample_after to determine per paper what is the largest sample size
after  ----
data_2 %<>%
  group_by(DOI) %>%
  mutate(largest_sample_after = if_else(is.na(DOI), -2,
                                        if_else(!is.na(final_sample_after),
as.numeric(max(final_sample_after,na.rm = TRUE)),
                                                -1
                                        )
  )
  )

data_2 %<>%
  group_by(Title) %>%
  mutate(largest_sample_after = if_else(!is.na(DOI), largest_sample_after,
                                        if_else(!is.na(final_sample_after),
as.numeric(max(final_sample_after,na.rm = TRUE)),
                                                -1
                                        )
  )
  )

summary(data_2$largest_sample_after) # No -2 values on largest_sample_size_after

length(which(is.na(data_2$largest_sample_after))) # 0 missing values on
largest_sample_after, because they are -1 instead of NA
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
```

```
length(which(is.na(data_2$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.infinite(data_2$largest_sample_after))) # 0 inf values on
largest_sample_after

data_2 %<>%
  mutate(largest_sample_after = na_if(largest_sample_after, "-1"))

length(which(is.na(data_2$largest_sample_after))) # 3 missing values on
largest_sample_after
length(which(is.na(data_2$DOI))) # 21 missing values on DOI
length(which(is.na(data_2$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.infinite(data_2$largest_sample_after))) # 0 inf values on
largest_sample_after

# Export the dataset to Excel ----
# path <- insert_your_own_path
path <- "C:\\Users\\Celess\\Documents\\Step02_MergedDataLargestSamples_V5.xlsx"
write_xlsx(data_2, path)
rm(path)
```

R Code for Adding Which Study Has the Largest Sample

This section contains the *R* code for determining which study within each paper has the largest sample size in said paper. For example, if a paper reports about one study, the study number 1 automatically is coded as the study number with the largest sample size within that paper.

```
# Empty the Global Environment ----
rm(list = ls())

# Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("writexl")

# Import the merged dataset from Excel ----
# path <- insert_your_own_path_to_the_file:
"Step02_MergedDataLargestSamples_V5.xlsx"
path <- "C:\\Users\\Celess\\Documents\\Step02_MergedDataLargestSamples_V5.xlsx"

data_1 <- read_excel(path,
                     sheet = 1,
                     guess_max = 21474836)

rm(path)

# Adding a new column for indicating per paper which study has the largest sample
size before exclusion----
data_1 %<>%
  mutate(largest_study_nr_before = NA,
         largest_study_nr_before = as.numeric(largest_study_nr_before)) # Create
empty column

data_1 <- data_1[order(data_1$DOI),] # Order data (ascending) on DOI

## Before exclusion ----
data_2 <- data_1
data_2 %<>%
  group_by(DOI) %>%
  mutate(largest_study_nr_before = if_else(is.na(largest_sample_before), -3, # -3
if largest_sample_before is missing
                                           if_else(is.na(final_sample_before), -4,
# -4 if final_sample_before is missing, but not largest_sample_before is not
missing
```

```
                                                    if_else(largest_sample_before ==
final_sample_before, study_number, # study_number if final_sample_before and
largest_sample_before are equal (and not missing)
                                                            0 # 0 if
final_sample_before and largest_sample_before are not equal (and not missing)
                                                    )
                                            )
  )
  )

# If largest_study_nr_before = 0, then that study is not the study with largest
sample size of that paper
# If largest_study_nr_before = 1, then that study is the study with largest sample
size of that paper
unique(data_2$largest_study_nr_before) # Values: 1 -3  3 0  2  4  5  6
View(data_2[, c("Title", "DOI", "final_sample_before", "largest_sample_before",
"study_number", "largest_study_nr_before")])

unique(data_2$largest_study_nr_before) # no -4 values on largest_study_nr_before

data_2 %<>%
  mutate(largest_study_nr_before = na_if(largest_study_nr_before, "-3"))

summary(data_2)
length(which(is.na(data_2$largest_study_nr_before))) # 240 missing values on
largest_study_nr_before
length(which(is.na(data_2$largest_sample_before))) # 240 missing values on
largest_sample_before
length(which(is.na(data_2$study_number))) # 0 missing values on study_number
length(which(is.na(data_2$decision_AV_SS_before))) # 1157 missing values on
decision_AV_SS_before
length(which(is.na(data_2$sample.size.before.exclusion))) # 267 missing values on
sample.size.before.exclusion

# Adding a new column for indicating per paper which study has the largest sample
size after exclusion----
data_3 <- data_2
data_3 %<>%
  mutate(largest_study_nr_after = NA,
         largest_study_nr_after = as.numeric(largest_study_nr_after) # Create empty
column
  )

data_3 <- data_3[order(data_3$DOI),] # Order data (ascending) on DOI

## After exclusion ----
data_2 <- data_3
data_2 %<>%
  group_by(DOI) %>%
  mutate(largest_study_nr_after = if_else(is.na(largest_sample_after), -3,
                                          if_else(is.na(final_sample_after), -4,
                                                  if_else(largest_sample_after ==
final_sample_after, study_number,
                                                          0
                                                  )
                                          )
  )
  )


# If largest_study_nr_after = 0, then that study is not the study with largest
sample size of that paper
# If largest_study_nr_after = 1, then that study is the study with largest sample
size of that paper
unique(data_2$largest_study_nr_after) # Values: 1  0  3  2  5  4  6  -3
View(data_2[, c("Title", "DOI", "final_sample_after", "largest_sample_after",
"study_number", "largest_study_nr_after")])
# No -4 values on largest_study_nr_before
```

```
data_2 %<>%
  mutate(largest_study_nr_after = na_if(largest_study_nr_after, "-3"))

data_3 <- data_2
summary(data_3)
length(which(is.na(data_3$largest_study_nr_after))) # 3 missing values on
largest_study_nr_after
length(which(is.na(data_3$largest_sample_after))) # 3 missing values on
largest_sample_after
length(which(is.na(data_3$study_number))) # 0 missing values on study_number
length(which(is.na(data_3$decision_AV_SS_after))) # 1158 missing values on
decision_AV_SS_after
length(which(is.na(data_3$sample_size_after.x))) # 23 missing values on
sample_size_after.x

# Export the dataset to Excel ----
# path <- insert_your_own_path
path <- "C:\\Users\\Celess\\Documents\\Step01_02_03_V7.xlsx"
write_xlsx(data_3, path)
rm(path)
```

R Code for Fixing One Specific Paper

This section contains the *R* code for manually fixing the incorrectly coded sample size of study 1 of the paper 'Social Judgment and Social Memory' from 50 to 1185.

```
#---- Empty the Global Environment ----
rm(list = ls())

#---- Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("writexl")

# Import the Excel file resulting from step 01, 02, and 03 ----
# path <- insert_your_own_path_to_the_file: "Step01_02_03_V7.xlsx""
path <- "C:\\Users\\Celess\\Documents\\Step01_02_03_V7.xlsx" # 1256 observations
and 228 variables

data_clean <- read_excel(path,
                         sheet = 1,
                         guess_max = 21474836)

rm(path)

# Fix the incorrect values of paper SOCIAL JUDGMENT AND SOCIAL MEMORY ----
data_fixed <- data_clean
# Changing the incorrectly coded sample size of study 1 ----
data_fixed %<>%
  mutate(sample.size.before.exclusion = ifelse(DOI != "10.1037/0022-3514.52.4.689",
sample.size.before.exclusion,
                                               ifelse(study_number == 1, 1185,

sample.size.before.exclusion)),
         sample_size_after.x = ifelse(DOI != "10.1037/0022-3514.52.4.689",
sample_size_after.x,
                                      ifelse(study_number == 1, 1185,
                                             sample_size_after.x)),
         coder_comment = ifelse(DOI != "10.1037/0022-3514.52.4.689", coder_comment,
                                ifelse(study_number == 1, "CD: sample size was
incorrectly coded as 50, but should be 1185",
                                       coder_comment)),
         final_sample_before = ifelse(DOI != "10.1037/0022-3514.52.4.689",
final_sample_before,
                                      ifelse(study_number == 1, 1185,
                                             final_sample_before)),
```

```
         final_sample_after = ifelse(DOI != "10.1037/0022-3514.52.4.689",
final_sample_after,
                                     ifelse(study_number == 1, 1185,
                                            final_sample_after)),
         largest_sample_before = ifelse(DOI != "10.1037/0022-3514.52.4.689",
largest_sample_before,
                                       ifelse(study_number == 1, 1185,
                                              largest_sample_before)),
         largest_sample_after = ifelse(DOI != "10.1037/0022-3514.52.4.689",
largest_sample_after,
                                      ifelse(study_number == 1, 1185,
                                             largest_sample_after)),
         largest_study_nr_before = ifelse(DOI != "10.1037/0022-3514.52.4.689",
largest_study_nr_before,
                                         ifelse(study_number == 1, 1,
                                                largest_study_nr_before)),
         largest_study_nr_after = ifelse(DOI != "10.1037/0022-3514.52.4.689",
largest_study_nr_after,
                                        ifelse(study_number == 1, 1,
                                               largest_study_nr_after))
  )

data_fixed %>%
  filter(DOI == "10.1037/0022-3514.52.4.689") %>%
  View()

# Adding study 2 including the right sample size ----
which(data_fixed$DOI == "10.1037/0022-3514.52.4.689", arr.ind=TRUE) # row number
520
data_fixed2 <- data_fixed %>%
  filter(row_number() %in% 520) %>%
  bind_rows(data_fixed)
which(data_fixed2$DOI == "10.1037/0022-3514.52.4.689", arr.ind=TRUE) # row number 1
and 521

data_fixed2[1, 220] <- 2 # change the study number (column 220) of paper Social
Judgment (row 1) to 2

data_fixed2 %<>%
  mutate(from_extra_data = ifelse(DOI != "10.1037/0022-3514.52.4.689",
from_extra_data,
                                  ifelse(study_number == 2, 1, #
Indicate that study number 2 is from the extra data
                                         from_extra_data)),
         sample.size.before.exclusion = ifelse(DOI != "10.1037/0022-3514.52.4.689",
sample.size.before.exclusion,
                                               ifelse(study_number == 2, 96,

sample.size.before.exclusion)),
         sample_size_after.x = ifelse(DOI != "10.1037/0022-3514.52.4.689",
sample_size_after.x,
                                      ifelse(study_number == 2, 96,
                                             sample_size_after.x)),
         coder_comment = ifelse(DOI != "10.1037/0022-3514.52.4.689", coder_comment,
                                ifelse(study_number == 2, "CD: study 2 was
initially not included in the data",
                                       coder_comment)),
         final_sample_before = ifelse(DOI != "10.1037/0022-3514.52.4.689",
final_sample_before,
                                      ifelse(study_number == 2, 96,
                                             final_sample_before)),
         final_sample_after = ifelse(DOI != "10.1037/0022-3514.52.4.689",
final_sample_after,
                                     ifelse(study_number == 2, 96,
                                            final_sample_after)),
         largest_study_nr_before = ifelse(DOI != "10.1037/0022-3514.52.4.689",
largest_study_nr_before,
                                          ifelse(study_number == 2, 0,
```

```
                                                largest_study_nr_before)),
        largest_study_nr_after = ifelse(DOI != "10.1037/0022-3514.52.4.689",
largest_study_nr_after,
                                        ifelse(study_number == 2, 0,
                                               largest_study_nr_after)),
        coder_sample_size_check = ifelse(DOI != "10.1037/0022-3514.52.4.689",
coder_sample_size_check,
                                         ifelse(study_number == 2, NA,
                                                coder_sample_size_check)),
        p_oa = ifelse(DOI != "10.1037/0022-3514.52.4.689", p_oa,
                      ifelse(study_number == 2, NA,
                             p_oa)),
        p_oa_gold = ifelse(DOI != "10.1037/0022-3514.52.4.689", p_oa_gold,
                           ifelse(study_number == 2, NA,
                                  p_oa_gold)),
        p_oa_bronze = ifelse(DOI != "10.1037/0022-3514.52.4.689", p_oa_bronze,
                             ifelse(study_number == 2, NA,
                                    p_oa_bronze)),
        p_oa_hybrid = ifelse(DOI != "10.1037/0022-3514.52.4.689", p_oa_hybrid,
                             ifelse(study_number == 2, NA,
                                    p_oa_hybrid)),
        p_oa_green = ifelse(DOI != "10.1037/0022-3514.52.4.689", p_oa_green,
                            ifelse(study_number == 2, NA,
                                   p_oa_green)),
        pub_block_begin = ifelse(DOI != "10.1037/0022-3514.52.4.689",
pub_block_begin,
                                 ifelse(study_number == 2, NA,
                                        pub_block_begin)),
        pub_block_end = ifelse(DOI != "10.1037/0022-3514.52.4.689", pub_block_end,
                               ifelse(study_number == 2, NA,
                                      pub_block_end)),
        p = ifelse(DOI != "10.1037/0022-3514.52.4.689", p,
                   ifelse(study_number == 2, NA,
                          p)),
        mcs = ifelse(DOI != "10.1037/0022-3514.52.4.689", mcs,
                     ifelse(study_number == 2, NA,
                            mcs)),
        tcs = ifelse(DOI != "10.1037/0022-3514.52.4.689", tcs,
                     ifelse(study_number == 2, NA,
                            tcs)),
        mncs = ifelse(DOI != "10.1037/0022-3514.52.4.689", mncs,
                      ifelse(study_number == 2, NA,
                             mncs)),
        mnjs = ifelse(DOI != "10.1037/0022-3514.52.4.689", mnjs,
                      ifelse(study_number == 2, NA,
                             mnjs)),
        pp_top_perc = ifelse(DOI != "10.1037/0022-3514.52.4.689", pp_top_perc,
                             ifelse(study_number == 2, NA,
                                    pp_top_perc)),
        pp_uncited = ifelse(DOI != "10.1037/0022-3514.52.4.689", pp_uncited,
                            ifelse(study_number == 2, NA,
                                   pp_uncited)),
        prop_self_cits = ifelse(DOI != "10.1037/0022-3514.52.4.689",
prop_self_cits,
                                ifelse(study_number == 2, NA,
                                       prop_self_cits)),
        int_cov = ifelse(DOI != "10.1037/0022-3514.52.4.689", int_cov,
                         ifelse(study_number == 2, NA,
                                int_cov)),
        pp_collab = ifelse(DOI != "10.1037/0022-3514.52.4.689", pp_collab,
                           ifelse(study_number == 2, NA,
                                  pp_collab)),
        pp_uncited = ifelse(DOI != "10.1037/0022-3514.52.4.689", pp_uncited,
                            ifelse(study_number == 2, NA,
                                   pp_uncited)),
        prop_self_cits = ifelse(DOI != "10.1037/0022-3514.52.4.689",
prop_self_cits,
                                ifelse(study_number == 2, NA,
```

```
                                        prop_self_cits)),
         int_cov = ifelse(DOI != "10.1037/0022-3514.52.4.689", int_cov,
                          ifelse(study_number == 2, NA,
                                 int_cov)),
         pp_collab = ifelse(DOI != "10.1037/0022-3514.52.4.689", pp_collab,
                            ifelse(study_number == 2, NA,
                                   pp_collab)),
         pp_int_collab = ifelse(DOI != "10.1037/0022-3514.52.4.689", pp_int_collab,
                                ifelse(study_number == 2, NA,
                                       pp_int_collab)),
         Z9 = ifelse(DOI != "10.1037/0022-3514.52.4.689", Z9,
                     ifelse(study_number == 2, NA,
                            Z9)),
         U1 = ifelse(DOI != "10.1037/0022-3514.52.4.689", U1,
                     ifelse(study_number == 2, NA,
                            U1)),
         U2 = ifelse(DOI != "10.1037/0022-3514.52.4.689", U2,
                     ifelse(study_number == 2, NA,
                            U2)),
         altmetric_jid = ifelse(DOI != "10.1037/0022-3514.52.4.689", altmetric_jid,
                                ifelse(study_number == 2, NA,
                                       altmetric_jid)),
         context.all.count = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.all.count,
                                    ifelse(study_number == 2, NA,
                                           context.all.count)),
         context.all.mean = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.all.mean,
                                   ifelse(study_number == 2, NA,
                                          context.all.mean)),
         context.all.rank = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.all.rank,
                                   ifelse(study_number == 2, NA,
                                          context.all.rank)),
         context.all.pct = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.all.pct,
                                  ifelse(study_number == 2, NA,
                                         context.all.pct)),
         context.all.higher_than = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.all.higher_than,
                                          ifelse(study_number == 2, NA,
                                                 context.all.higher_than)),
         context.similar_age_3m.count = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.similar_age_3m.count,
                                               ifelse(study_number == 2, NA,

context.similar_age_3m.count)),
         context.similar_age_3m.mean = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.similar_age_3m.mean,
                                              ifelse(study_number == 2, NA,
                                                     context.similar_age_3m.mean)),
         context.similar_age_3m.rank = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.similar_age_3m.rank,
                                              ifelse(study_number == 2, NA,
                                                     context.similar_age_3m.rank)),
         context.similar_age_3m.pct = ifelse(DOI != "10.1037/0022-3514.52.4.689",
context.similar_age_3m.pct,
                                             ifelse(study_number == 2, NA,
                                                    context.similar_age_3m.pct)),
         context.similar_age_3m.higher_than = ifelse(DOI != "10.1037/0022-
3514.52.4.689", context.similar_age_3m.higher_than,
                                                     ifelse(study_number == 2, NA,

context.similar_age_3m.higher_than))
  )

data_fixed2 %>%
  filter(DOI == "10.1037/0022-3514.52.4.689") %>%
```

```
  View()

# Export the file with fixed sample sizes for Social Judgment ----
# Insert your own path to the file
path <- "C:\\Users\\celes\\Documents\\Thesis 2021\\Step01-04_V10.xlsx"
write_xlsx(data_fixed2, path)
rm(path)
```

R Code for Completing Citation Scores

This section contains the *R* code for completing missing citation scores of the extra studies after manually looking the corresponding papers up in the master file.

```
#---- Empty the Global Environment ----
rm(list = ls())

#---- Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("writexl")

# Import the Excel file resulting from step 01, 02, 03, and 04 ----
# path <- insert_your_own_path_to_the_file: "Step01-04_V10.xlsx""
path <- "C:\\Users\\Celess\\Documents\\Step01-04_V10.xlsx" # 1257 observations and
228 variables

data_clean <- read_excel(path,
                         sheet = 1,
                         guess_max = 21474836)

rm(path)

# Fill in the missing values of Total Citation Score for the extra studies ----
length(which(is.na(data_clean$study_number))) # 0
length(which(is.na(data_clean$TC))) # 297
length(which(is.na(data_clean$DOI))) # 21

data_1 <- data_clean
data_1 %<>%
  arrange(DOI) %>%
  group_by(DOI) %>%
  mutate(TC = if_else(!is.na(TC), TC,
                  if_else(study_number == 1, TC,
                  if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                       if_else(DOI == lag(DOI), # For the extra studies
                               lag(TC), -3
                       )
                  )
              )
  ),
  TC = if_else(!is.na(TC), TC,
               if_else(study_number == 1, TC,
                        if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                                if_else(DOI == lag(DOI), # For the extra studies
                                        lag(TC), -3
                                )
                        )
               )
  ),
  TC = if_else(!is.na(TC), TC,
               if_else(study_number == 1, TC,
                        if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                                if_else(DOI == lag(DOI), # For the extra studies
                                        lag(TC), -3
```

```
                               )
                       )
               )
  ),
  TC = if_else(!is.na(TC), TC,
               if_else(study_number == 1, TC,
                       if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                               if_else(DOI == lag(DOI), # For the extra studies
                                       lag(TC), -3
                               )
                       )
               )
  ),
  TC = if_else(!is.na(TC), TC,
               if_else(study_number == 1, TC,
                       if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                               if_else(DOI == lag(DOI), # For the extra studies
                                       lag(TC), -3
                               )
                       )
               )
  ),
  TC = if_else(!is.na(TC), TC,
               if_else(study_number == 1, TC,
                       if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                               if_else(DOI == lag(DOI), # For the extra studies
                                       lag(TC), -3
                               )
                       )
               )
  ),
  TC = if_else(!is.na(TC), TC,
               if_else(study_number == 1, TC,
                       if_else(is.na(DOI), -2, # TC changes from NA to -2 if DOI is
missing and study_number is not 1
                               if_else(DOI == lag(DOI), # For the extra studies
                                       lag(TC), -3
                               )
                       )
               )
  )
  )

summary(data_1$TC) # min = -2 (thus no -3)
length(which(data_1$TC == -2)) # 2 extra studies with missing DOI: "FEELINGS OF
MASTERY IN HIGH AGGRESSION-HISTORY AGGRESSORS" and "FUTURE-ORIENTED PEOPLE SHOW
STRONGER MORAL CONCERNS"
length(which(is.na(data_1$study_number))) # 0
length(which(is.na(data_1$TC))) # 40
length(which(is.na(data_1$DOI))) # 21

      # data_1 %>%
      #   select(Title, DOI, TC, study_number, final_sample_after,
largest_sample_after, largest_study_nr_after) %>%
      #   View()

data_2 <- data_1 %>%
  mutate(TC = na_if(TC, "-2"))

summary(data_2$TC) # min = 0 (thus no -2)
length(which(data_2$TC == -2)) # 0
length(which(is.na(data_2$study_number))) # 0
length(which(is.na(data_2$TC))) # 42
length(which(is.na(data_2$DOI))) # 21
length(which(is.infinite(data_2$TC))) # 0 inf values on TC
```

```
# Export the file with filled in citation scores for the extra studies ----
# Insert your own path to the file
path <- "C:\\Users\\celes\\Documents\\Thesis 2021\\Step01-04_V11.xlsx"
write_xlsx(data_2, path)
rm(path)
```

R Code for Study Numbers, Exclusions and Calculating RV

This section contains the *R* code for describing the study numbers, applying the exclusion criteria on the cleaned dataset and then calculating RV for all studies in the final dataset.

```
#---- Empty the Global Environment ----
rm(list = ls())

#---- Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("Hmisc") # for using describe()
library("writexl")
library(fmsb) # for radarchart()
library("formattable") # for formattable()

#---- Apa theme for plots ----
# Ref: https://rdrr.io/cran/jtools/src/R/theme_apa.R
add_gridlines <- function(x = TRUE, y = TRUE, minor = TRUE) {

  plot <- theme()

  if (y == TRUE) {
    plot <- plot + theme(panel.grid.major.y = element_line(colour = "grey92"))
    if (minor == TRUE) {
      plot <-
        plot + theme(panel.grid.minor.y = element_line(colour = "grey92",
                                                       size = .25))
    }
  }

  if (x == TRUE) {
    plot <- plot + theme(panel.grid.major.x = element_line(colour = "grey92"))
    if (minor == TRUE) {
      plot <-
        plot + theme(panel.grid.minor.x = element_line(colour = "grey92",
                                                       size = .25))
    }
  }

  return(plot)

}

add_x_gridlines <- function(minor = TRUE) {
  add_gridlines(x = TRUE, y = FALSE, minor = minor)
}

add_y_gridlines <- function(minor = TRUE) {
  add_gridlines(x = FALSE, y = TRUE, minor = minor)
}

drop_gridlines <- function(x = TRUE, y = TRUE, minor.only = FALSE) {

  plot <- ggplot2::theme()

  if (y == TRUE) {
    plot <- plot + ggplot2::theme(panel.grid.minor.y = ggplot2::element_blank())
```

```
    if (minor.only == FALSE) {
      plot <-
        plot + ggplot2::theme(panel.grid.major.y = ggplot2::element_blank())
    }
  }

  if (x == TRUE) {
    plot <- plot + ggplot2::theme(panel.grid.minor.x = ggplot2::element_blank())
    if (minor.only == FALSE) {
      plot <-
        plot + ggplot2::theme(panel.grid.major.x = ggplot2::element_blank())
    }
  }

  return(plot)

}

drop_x_gridlines <- function(minor.only = FALSE) {
  drop_gridlines(x = TRUE, y = FALSE, minor.only = minor.only)
}

drop_y_gridlines <- function(minor.only = FALSE) {
  drop_gridlines(x = FALSE, y = TRUE, minor.only = minor.only)
}

theme_apa <- function(
  legend.pos = "right",
  legend.use.title = FALSE,
  legend.font.size = 12,
  x.font.size = 12,
  y.font.size = 12,
  facet.title.size = 12,
  remove.y.gridlines = TRUE,
  remove.x.gridlines = TRUE
) {

  # Specifying parameters, using theme_bw() as starting point
  plot <- ggplot2::theme_bw() + ggplot2::theme(
    plot.title = ggplot2::element_text(face = "bold", hjust = 0, size = 14),
    axis.title.x = ggplot2::element_text(size = x.font.size),
    axis.title.y = ggplot2::element_text(size = y.font.size,
                                         angle = 90),
    legend.text = ggplot2::element_text(size = legend.font.size),
    legend.key.size = ggplot2::unit(1.5, "lines"),
    # switch off the rectangle around symbols
    legend.key = ggplot2::element_blank(),
    legend.key.width = grid::unit(2, "lines"),
    strip.text.x = ggplot2::element_text(size = facet.title.size), # facet labs
    strip.text.y = ggplot2::element_text(size = facet.title.size),
    # facet titles
    strip.background = ggplot2::element_rect(colour = "white", fill = "white"),
    panel.background = ggplot2::element_rect(fill = "white"),
    plot.title.position = "panel",
    # complete = TRUE
  )

  # Choose legend position. APA figures generally include legends that
  # are embedded on the plane, so there is no efficient way to have it
  # automatically placed correctly
  if (legend.pos == "topleft") {
    # manually position the legend (numbers being from 0,0 at bottom left of
    # whole plot to 1,1 at top right)
    plot <- plot + ggplot2::theme(legend.position = c(.05, .95),
                                  legend.justification = c(.05, .95))
  } else if (legend.pos == "topright") {
    plot <- plot + ggplot2::theme(legend.position = c(.95, .95),
                                  legend.justification = c(.95, .95))
```

```
  } else if (legend.pos == "topmiddle") {
    plot <- plot + ggplot2::theme(legend.position = c(.50, .95),
                                  legend.justification = c(.50, .95))
  } else if (legend.pos == "bottomleft") {
    plot <- plot + ggplot2::theme(legend.position = c(.05, .05),
                                  legend.justification = c(.05, .05))
  } else if (legend.pos == "bottomright") {
    plot <- plot + ggplot2::theme(legend.position = c(.95, .05),
                                  legend.justification = c(.95, .05))
  } else if (legend.pos == "bottommiddle") {
    plot <- plot + ggplot2::theme(legend.position = c(.50, .05),
                                  legend.justification = c(.50, .05))
  } else if (legend.pos == "none") {
    plot <- plot + ggplot2::theme(legend.position = "none")
  } else {
    plot <- plot + ggplot2::theme(legend.position = legend.pos)
  }

  # Should legend have title? If so, format it correctly
  if (legend.use.title == FALSE) {
    # switch off the legend title
    plot <- plot +
      ggplot2::theme(legend.title = ggplot2::element_blank())

  } else {
    plot <- plot +
      ggplot2::theme(legend.title =
                       ggplot2::element_text(size = 12, face = "bold"))
  }

  if (remove.y.gridlines == TRUE) {
    plot <- plot + drop_y_gridlines()
  } else {
    plot <- plot + add_y_gridlines()
  }

  if (remove.x.gridlines == TRUE) {
    plot <- plot + drop_x_gridlines()
  } else {
    plot <- plot + add_x_gridlines()
  }

  return(plot)

}

# Import the clean data ----
# path <- insert_your_own_path_to_the_file: "Step01-04_V11.xlsx"
path <- "C:\\Users\\celes\\Documents\\Thesis 2021\\Step01-04_V11.xlsx" # 1257
observations and 228 variables

data_clean <- read_excel(path,
                         sheet = 1,
                         guess_max = 21474836)

rm(path)

#---- Describe frequencies of study numbers ----
range(data_clean$study_number) # 1-6
describe(data_clean$study_number)
length(which(is.na(data_clean$study_number))) # 0
length(which(data_clean$study_number == 1)) # 999
length(which(data_clean$study_number == 2)) # 173
length(which(data_clean$study_number == 3)) # 60
length(which(data_clean$study_number == 4)) # 16
length(which(data_clean$study_number == 5)) # 6
length(which(data_clean$study_number == 6)) # 3
# 999+173+60+16+6+3 = 1257
```

```
# Number of papers reporting 6 studies: 3
# Number of papers reporting 5 studies: 6 - 3 = 3
# Number of papers reporting 4 studies: 16 - 6 - 3 = 7
# Number of papers reporting 3 studies: 60 - 16 - 6 - 3 = 35
# Number of papers reporting 2 studies: 173 - 60 - 16 - 6 - 3 = 88
# Number of papers reporting 1 studies: 999 - 173 - 60 - 16 - 6 - 3 = 741
# 3 + 3 + 7 + 35 + 88 + 741 = 877

#---- Exclude papers with missing DOI ----
describe(data_clean$DOI) # 21 missings; 980 distinct values
data_excl1 <- data_clean %>%
  filter(!is.na(DOI))
length(which(is.na(data_excl1$DOI))) # 0 observations with unknown DOI
# 1257 observations in data_clean - 21 missing DOI = 1236 observations in
data_excl1

#---- Exclude papers published later than 2018 ----
describe(data_excl1$Publication.Year) # 0 missings; 66 distinct values
length(which(data_excl1$Publication.Year > 2018)) # 34 observations published later
than 2018
data_excl2 <- data_excl1 %>%
  filter(Publication.Year <= 2018)
length(which(data_excl2$Publication.Year > 2018)) # 0 observations published later
than 2018
# 1236 observations in data_excl1 - 34 too recent = 1202 observations in data_excl2

#---- Exclude papers with missing Total Citation Score ----
describe(data_clean$TC) # 42 missings; 135 distinct values
describe(data_excl2$TC) # 21 missings; 135 distinct values
data_excl3 <- data_excl2 %>%
  filter(!is.na(TC))
describe(data_excl3$TC) # 0 missings; 135 distinct values
# 1202 observations in data_excl2 - 21 missing TC = 1181 observations in data_excl3

#---- Exclude papers with missing sample size ----
describe(data_clean$final_sample_after) # 21 missings; 475 distinct values
describe(data_excl3$final_sample_after) # 0 missings; 459 distinct values
data_excl4 <- data_excl3 %>%
  filter(!is.na(final_sample_after))
# 1181 observations in data_excl3 - 0 missing final_sample_after = 1181
observations in data_excl4

#---- Exclude papers with missing publication year ----
describe(data_clean$Publication.Year) # 0 missings; 66 distinct values
describe(data_excl3$Publication.Year) # 0 missings; 65 distinct values
data_excl5 <- data_excl4 %>%
  filter(!is.na(Publication.Year))
# 1181 observations in data_excl4 - 0 missing Publication.Year = 1181 observations
in data_excl5

# Horizontal barplot of journals with >=10 articles ----
describe(data_excl5$JI) # n = 923; 258 missings; 64 distinct journals
overview_ji <- as.data.frame(table(data_excl5$JI))
overview_ji[order(overview_ji$Freq, decreasing = TRUE), ] # "PERS. INDIVID.
DIFFER." has highest frequency (freq = 99)
#The journal most articles were published in was 'Personality and Individual
Differences'(99 times)
#   J. PERS. SOC. PSYCHOL.   87
#   J. SOC. PSYCHOL.   69

subset_ji <- subset(overview_ji, Freq >= 10) # Select only journals with >= 10
articles
subset_ji[order(subset_ji$Freq, decreasing = TRUE), ]

ggplot(subset_ji, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of journals with at least 10 articles") + theme(plot.title =
element_text(hjust = 0.5)) +
```

```
  xlab("Journal") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,100), expand = c(0, 0)) +
  geom_text(label=subset_ji$Freq, vjust = -0.4, hjust = 0, size = 3,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 8, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50)) +
  coord_flip()

#---- Exclude studies that do not have the largest sample size within a paper ----
describe(data_clean$largest_study_nr_after) # 21 missings; 7 distinct values
describe(data_excl5$largest_study_nr_after) # 0 missings; 7 distinct values
# If largest_study_nr_after == 0, then that row/observation/study does not have the
largest sample size
# If largest_study_nr_after != 0, then that row/observation/study has the largest
sample size

        # data_excl5 %>%
        #   select(Title, DOI, TC, study_number, final_sample_after,
largest_sample_after, largest_study_nr_after) %>%
        #   View()

data_excl6 <- data_excl5 %>%
  filter(largest_study_nr_after != 0)
# 1181 observations in data_excl5 - 244 non-zero largest_study_nr_after = 937
observations in data_excl6

data_final <- data_excl6

#---- Calculate RV for all records in data_final ----
current_year <- 2020 # Used to determine how old papers are compared to the current
year

data_finalRV <- data_final %>%
  mutate(years_since_pub = current_year - Publication.Year,
         RV = TC / (years_since_pub + 1) * (1 / final_sample_after), # is
automatically the largest sample after, because of exclusion 6
  )

# Export the dataset to Excel ----
# path <- insert_your_own_path_to_the_file: "Step01-04_V12.xlsx"
path <- "C:\\Users\\celes\\Documents\\Thesis 2021\\Step01-04_V12.xlsx"
write_xlsx(data_finalRV, path)
rm(path)
```

R Code for Sample Descriptives and Extracting Top, Center, and Bottom 10

This section contains the *R* code for generating plots that describe the sample and then for obtaining the top, center, and bottom 10 studies.

```
#---- Empty the Global Environment ----
rm(list = ls())

#---- Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("Hmisc") # for using describe()
library("writexl")
library(fmsb) # for radarchart()
library("formattable") # for formattable()

#---- Apa theme for plots ----
# Ref: https://rdrr.io/cran/jtools/src/R/theme_apa.R
add_gridlines <- function(x = TRUE, y = TRUE, minor = TRUE) {
```

```
  plot <- theme()

  if (y == TRUE) {
    plot <- plot + theme(panel.grid.major.y = element_line(colour = "grey92"))
    if (minor == TRUE) {
      plot <-
        plot + theme(panel.grid.minor.y = element_line(colour = "grey92",
                                                       size = .25))
    }
  }

  if (x == TRUE) {
    plot <- plot + theme(panel.grid.major.x = element_line(colour = "grey92"))
    if (minor == TRUE) {
      plot <-
        plot + theme(panel.grid.minor.x = element_line(colour = "grey92",
                                                       size = .25))
    }
  }

  return(plot)

}

add_x_gridlines <- function(minor = TRUE) {
  add_gridlines(x = TRUE, y = FALSE, minor = minor)
}

add_y_gridlines <- function(minor = TRUE) {
  add_gridlines(x = FALSE, y = TRUE, minor = minor)
}

drop_gridlines <- function(x = TRUE, y = TRUE, minor.only = FALSE) {

  plot <- ggplot2::theme()

  if (y == TRUE) {
    plot <- plot + ggplot2::theme(panel.grid.minor.y = ggplot2::element_blank())
    if (minor.only == FALSE) {
      plot <-
        plot + ggplot2::theme(panel.grid.major.y = ggplot2::element_blank())
    }
  }

  if (x == TRUE) {
    plot <- plot + ggplot2::theme(panel.grid.minor.x = ggplot2::element_blank())
    if (minor.only == FALSE) {
      plot <-
        plot + ggplot2::theme(panel.grid.major.x = ggplot2::element_blank())
    }
  }

  return(plot)

}

drop_x_gridlines <- function(minor.only = FALSE) {
  drop_gridlines(x = TRUE, y = FALSE, minor.only = minor.only)
}

drop_y_gridlines <- function(minor.only = FALSE) {
  drop_gridlines(x = FALSE, y = TRUE, minor.only = minor.only)
}

theme_apa <- function(
  legend.pos = "right",
  legend.use.title = FALSE,
```

```
  legend.font.size = 12,
  x.font.size = 12,
  y.font.size = 12,
  facet.title.size = 12,
  remove.y.gridlines = TRUE,
  remove.x.gridlines = TRUE
) {

  # Specifying parameters, using theme_bw() as starting point
  plot <- ggplot2::theme_bw() + ggplot2::theme(
    plot.title = ggplot2::element_text(face = "bold", hjust = 0, size = 14),
    axis.title.x = ggplot2::element_text(size = x.font.size),
    axis.title.y = ggplot2::element_text(size = y.font.size,
                                         angle = 90),
    legend.text = ggplot2::element_text(size = legend.font.size),
    legend.key.size = ggplot2::unit(1.5, "lines"),
    # switch off the rectangle around symbols
    legend.key = ggplot2::element_blank(),
    legend.key.width = grid::unit(2, "lines"),
    strip.text.x = ggplot2::element_text(size = facet.title.size), # facet labs
    strip.text.y = ggplot2::element_text(size = facet.title.size),
    # facet titles
    strip.background = ggplot2::element_rect(colour = "white", fill = "white"),
    panel.background = ggplot2::element_rect(fill = "white"),
    plot.title.position = "panel",
    # complete = TRUE
  )

  # Choose legend position. APA figures generally include legends that
  # are embedded on the plane, so there is no efficient way to have it
  # automatically placed correctly
  if (legend.pos == "topleft") {
    # manually position the legend (numbers being from 0,0 at bottom left of
    # whole plot to 1,1 at top right)
    plot <- plot + ggplot2::theme(legend.position = c(.05, .95),
                                  legend.justification = c(.05, .95))
  } else if (legend.pos == "topright") {
    plot <- plot + ggplot2::theme(legend.position = c(.95, .95),
                                  legend.justification = c(.95, .95))
  } else if (legend.pos == "topmiddle") {
    plot <- plot + ggplot2::theme(legend.position = c(.50, .95),
                                  legend.justification = c(.50, .95))
  } else if (legend.pos == "bottomleft") {
    plot <- plot + ggplot2::theme(legend.position = c(.05, .05),
                                  legend.justification = c(.05, .05))
  } else if (legend.pos == "bottomright") {
    plot <- plot + ggplot2::theme(legend.position = c(.95, .05),
                                  legend.justification = c(.95, .05))
  } else if (legend.pos == "bottommiddle") {
    plot <- plot + ggplot2::theme(legend.position = c(.50, .05),
                                  legend.justification = c(.50, .05))
  } else if (legend.pos == "none") {
    plot <- plot + ggplot2::theme(legend.position = "none")
  } else {
    plot <- plot + ggplot2::theme(legend.position = legend.pos)
  }

  # Should legend have title? If so, format it correctly
  if (legend.use.title == FALSE) {
    # switch off the legend title
    plot <- plot +
      ggplot2::theme(legend.title = ggplot2::element_blank())

  } else {
    plot <- plot +
      ggplot2::theme(legend.title =
                       ggplot2::element_text(size = 12, face = "bold"))
  }
```

```
  if (remove.y.gridlines == TRUE) {
    plot <- plot + drop_y_gridlines()
  } else {
    plot <- plot + add_y_gridlines()
  }

  if (remove.x.gridlines == TRUE) {
    plot <- plot + drop_x_gridlines()
  } else {
    plot <- plot + add_x_gridlines()
  }

  return(plot)

}


#---- Importing the clean data  ---------------------------------------------------
-------------------------------------------------------------
# Import the Excel file resulting from: step 01, 02, 03, and FixingSocialJudgment
and filling in TC ----
# path <- insert_your_own_path_to_the_file: "Step01-04_V12.xlsx"
path <- "C:\\Users\\celes\\Documents\\Thesis 2021\\Step01-04_V12.xlsx" # 937
observations and 230 variables

data_clean <- read_excel(path,
                         sheet = 1,
                         guess_max = 21474836)

rm(path)

# Describing the clean data (incl. and excl. extra data) ----
describe(data_clean$from_extra_data) # 0 missings; 2 distinct values (0 = not from
extra data; 1 = from extra data)
describe(data_clean$DOI) # 0 missings; 931 distinct values
describe(data_clean$study_number) # 0 missings; 6 distinct values

data_extra <- data_clean %>% # 90 observations from extra data
  filter(from_extra_data == 1)
length(which(data_clean$study_number != 1)) # 90

data_not_extra <- data_clean %>% # 847 observations from master file (i.e., not
from extra data)
  filter(from_extra_data == 0)
length(which(data_clean$study_number == 1)) # 847

# Clean data
describe(data_clean$ID) # No missings; 931 distinct IDs
range(data_clean$ID) # 1-1000

# Vertical barplot (with angled x-axis labels) of all journal frequencies ----
describe(data_clean$JI) # n = 838; 99 missings; 64 distinct journals
overview_ji <- as.data.frame(table(data_clean$JI))
overview_ji[order(overview_ji$Freq, decreasing = TRUE), ] # "PERS. INDIVID.
DIFFER." has highest frequency (freq = 96)
#The journal most articles were published in was 'Personality and Individual
Differences'(96 times)

ggplot(overview_ji, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of all journals") + theme(plot.title = element_text(hjust =
0.5)) +
  xlab("Journal") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,100), expand = c(0, 0)) +
  geom_text(label=overview_ji$Freq, vjust = -0.4, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
```

```
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 45, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

# Horizontal barplot of journals with >=10 articles ----
subset_ji <- subset(overview_ji, Freq >= 10) # Select only journals with >10
articles
subset_ji[order(subset_ji$Freq, decreasing = TRUE), ]

ggplot(subset_ji, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of journals with at least 10 articles") + theme(plot.title =
element_text(hjust = 0.5)) +
  xlab("Journal") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,100), expand = c(0, 0)) +
  geom_text(label=subset_ji$Freq, vjust = -0.4, hjust = -1.2, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 8, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50)) +
  coord_flip()

# Vertical barplot of all publication year frequencies ----
describe(data_clean$Publication.Year) # 0 missings; 65 distinct publication years
overview_py <- as.data.frame(table(data_clean$Publication.Year))
overview_py[order(overview_py$Freq, decreasing = TRUE), ] # "2010" has highest
frequency (freq = 41)
#The year most articles were published in was 2010 (41 times)

ggplot(overview_py, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of all publication years") + theme(plot.title =
element_text(hjust = 0.5)) +
  xlab("Publication year") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,45), expand = c(0, 0)) +
  geom_text(label=overview_py$Freq, vjust = -0.4, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

# Vertical barplot (with angled x-axis labels) of all citation score frequencies --
--
describe(data_clean$TC) # 0 missings; 135 distinct citation scores
overview_tc <- as.data.frame(table(data_clean$TC))
overview_tc[order(overview_tc$Freq, decreasing = TRUE), ] # "1" has highest
frequency (freq = 52)
#Most articles (52) were cited 1 time

ggplot(overview_tc, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of all citation scores") + theme(plot.title =
element_text(hjust = 0.5)) +
  xlab("Citation score") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,60), expand = c(0, 0)) +
  geom_text(label=overview_tc$Freq, vjust = -0.4, size = 3,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 20, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

# Vertical barplot citation score with >= 2 frequencies ----
describe(data_clean$TC) # 0 missings; 135 distinct citation scores
subset_tc <- subset(overview_tc, Freq >= 2) # Select only citation scores with >= 2
frequencies
```

```
subset_tc[order(subset_tc$Freq, decreasing = TRUE), ]

plot.new()
ggplot(subset_tc, aes(x=Var1, y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of citation scores with frequency >= 2") + theme(plot.title =
element_text(hjust = 0.5)) +
  xlab("Citation score") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,60), expand = c(0, 0)) +
  geom_text(label=subset_tc$Freq, vjust = -0.4, size = 3,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  #theme(axis.text.x = element_text(size = 5, angle = 45, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50)) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 7, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

# Create new data set for the bottom 10 ----
length(which(data_clean$TC == 0)) # 50
# Select the 50 papers with TC == 0 and order them on final_sample_after
data_bottom10 <- data_clean %>%
  filter(TC == 0)

data_bottom10 <- data_bottom10[order(data_bottom10$final_sample_after),] # Order
data ascending on sample size
data_bottom10 <- tail(data_bottom10, n = 10)

data_bottom10 %>%
  select(ID, Authors, Title, Publication.Year, DOI, TC, final_sample_after,
study_number, RV) %>%
  View()

# Create new data set for the center 10 ----
# data_clean has 937 papers
937 / 2 # 468.5 rounded up is 469
469 - 4 # 465
469 + 5 # 474

data_center10 <- data_clean[order(-data_clean$RV),] # Order data descending on RVs
data_center10 <- data_center10[465:474, ]
data_center10 %>%
  select(ID, Authors, Title, Publication.Year, DOI, TC, final_sample_after,
study_number, RV) %>%
  View()

# Create new data set for the top 10 ----
data_top10 <- data_clean[order(-data_clean$RV),] # Order data descending on RVs
data_top10 <- head(data_top10, n = 10)

data_top10 %>%
  select(ID, Authors, Title, Publication.Year, DOI, TC, final_sample_after,
study_number, RV) %>%
  View()

# Horizontal barplot of all study numbers ----
describe(data_clean$study_number) # n = 937; 0 missings; 6 distinct study numbers
overview_studynr <- as.data.frame(table(data_clean$study_number))
overview_studynr[order(overview_studynr$Freq, decreasing = TRUE), ] # "1" has
highest frequency (freq = 847)

ggplot(overview_studynr, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of all study numbers") + theme(plot.title = element_text(hjust
= 0.5)) +
  xlab("Study number") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,900), expand = c(0, 0)) +
```

```
  geom_text(label=overview_studynr$Freq, vjust = -0.4, hjust = -1.5, size =
3,position = position_dodge(width = 0.9),inherit.aes = TRUE) +
  #theme(axis.text.x = element_text(size = 5, angle = 45, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50)) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 8, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50)) +
  coord_flip()

# Vertical barplot (with angled x-axis labels) of all sample sizes ----
describe(data_clean$final_sample_after) # n = 937; 0 missings; 441 distinct values
overview_ss <- as.data.frame(table(data_clean$final_sample_after))
overview_ss[order(overview_ss$Freq, decreasing = TRUE), ] # Sample size of 40 has
highest frequency (freq = 21)

ggplot(overview_ss, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of all sample sizes") + theme(plot.title = element_text(hjust
= 0.5)) +
  xlab("Sample size") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,25), expand = c(0, 0)) +
  geom_text(label=overview_ss$Freq, vjust = -0.4, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 45, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

# Vertical barplot of sample sizes with >= 4 frequencies ----
subset_ss <- subset(overview_ss, Freq >= 4) # Select only sample sizes with >= 4
frequencies
subset_ss[order(subset_ss$Freq, decreasing = TRUE), ]

ggplot(subset_ss, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of sample sizes with frequency >= 4") + theme(plot.title =
element_text(hjust = 0.5)) +
  xlab("Sample size") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,25), expand = c(0, 0)) +
  geom_text(label=subset_ss$Freq, vjust = -0.4, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

# Vertical barplot (with angled x-axis labels) of Replication Values < 50
frequencies ----
describe(data_clean$RV) # n = 937; 0 missings; 851 distinct values
overview_RV <- as.data.frame(table(data_clean$RV))
overview_RV[order(overview_RV$Freq, decreasing = TRUE), ] # RV of 0 has highest
frequency (freq = 50)
subset_RV <- subset(overview_RV, Freq < 50) # Select only RVs with < 50 frequencies
subset_RV[order(subset_RV$Freq, decreasing = TRUE), ]

ggplot(subset_RV, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of Replication Values with frequency < 50") + theme(plot.title
= element_text(hjust = 0.5)) +
  xlab("Replication Value") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,3), expand = c(0, 0)) +
  geom_text(label=subset_RV$Freq, vjust = -0.4, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))
```

```
# Vertical barplot (with angled x-axis labels) of Replication Values > 0.001
frequencies ----
subset_RV2 <- subset(overview_RV, Freq > 0.001) # Select only RVs with > 0.001
frequencies
subset_RV2[order(subset_RV2$Freq, decreasing = TRUE), ]

overview_RV2 <- as.data.frame(table(log(data_clean$RV)))
overview_RV2[order(overview_RV2$Freq, decreasing = TRUE), ] # RV of 0 has highest
frequency (freq = 50)
subset_RV3 <- subset(overview_RV2, Freq < 50) # Select only RVs with < 50
frequencies
subset_RV3[order(subset_RV3$Freq, decreasing = TRUE), ]

ggplot(subset_RV3, aes(x=reorder(Var1, -Freq), y=Freq)) +
  geom_bar(stat = "Identity") +
  ggtitle("Frequency of Replication Values with frequency X") + theme(plot.title =
element_text(hjust = 0.5)) +
  xlab("Replication Value") + ylab("Frequency") +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_continuous(limits = c(0,50), expand = c(0, 0)) +
  geom_text(label=subset_RV3$Freq, vjust = -0.4, size = 3.8,position =
position_dodge(width = 0.9),inherit.aes = TRUE) +
  theme_apa() +
  theme(axis.text.x = element_text(size = 5, angle = 0, vjust = 1, hjust = 1),
plot.margin = margin(10, 20, 5, 50))

hist(data_clean$RV)
hist(log(data_clean$RV))
```

R Code for Radar Plots

This section contains the *R* code for generating the radar plots for the top, center, and bottom 10 studies.

```
#---- Empty the Global Environment ----
rm(list = ls())

#---- Libraries being used in this code ----
library("tidyverse")
library("readxl")
library("magrittr") # for using %<>%
library("writexl")
library(fmsb) # for radarchart()

# Import the excel file resulting from step 01, 02, 03 and 04 ----
# path <- insert_your_own_path_to_the_file: "Step01-04_V12.xlsx"
path <- "C:\\Users\\celes\\Documents\\Thesis 2021\\Step01-04_V12.xlsx" # 937
observations and 230 variables

data_clean <- read_excel(path,
                         sheet = 1,
                         guess_max = 21474836)

rm(path)

# Radar plots ----
# For paper 1 of the top 10 ----
data_1top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 1),
                          Sample = c(3, 0, 2),
                          Openness = c(3, 0, 3),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 3),
                          Pvalues = c(3, 0, 0)
)
```

```
plot.new()
radarchart(data_1top10,
           seg = 3,  # Number of axis segments
           title = "1: Testosterone and Chess Competition",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 2 of the top 10 ----
data_2top10 <- data.frame(Hypotheses = c(3, 0, 3), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 0),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 1),
                          Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_2top10,
           seg = 3,  # Number of axis segments
           title = "2: Generality of the Automatic Attitude Activation Effect",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 3 of the top 10 ----
data_3top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 0),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 3),
                          Pvalues = c(3, 0, 3)
)

plot.new()
radarchart(data_3top10,
           seg = 3,  # Number of axis segments
           title = "3: The Scrooge Effect: Evidence That Mortality Salience
Increases Prosocial Attitudes and Behavior",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 4 of the top 10 ----
data_4top10 <- data.frame(Hypotheses = c(3, 0, 3), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 3),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 2),
                          Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_4top10,
           seg = 3,  # Number of axis segments
           title = "4: Varieties of Disgust Faces and the Structure of Disgust",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 5 of the top 10 ----
data_5top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
```

```
                          Exclusion = c(3, 0, 1),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 3),
                          Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_5top10,
           seg = 3,  # Number of axis segments
           title = "5: Explaining the Enigmatic Anchoring Effect: Mechanisms of
Selective Accessibility",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 6 of the top 10 ----
data_6top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 1),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 3),
                          Pvalues = c(3, 0, 2)
)

plot.new()
radarchart(data_6top10,
           seg = 3,  # Number of axis segments
           title = "6: Attention in Delay of Gratification",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 7 of the top 10 ----
data_7top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 1),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 3),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 2),
                          Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_7top10,
           seg = 3,  # Number of axis segments
           title = "7: Is Empathy-Induced Helping Due to Self-Other Merging",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 8 of the top 10 ----
data_8top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 0),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 3),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 0),
                          Pvalues = c(3, 0, 3)
)

plot.new()
```

```
radarchart(data_8top10,
           seg = 3,  # Number of axis segments
           title = "8: When Approach Motivation and Behavioral Inhibition Collide:
Behavior Regulation Through Stimulus Devaluation",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 9 of the top 10 ----
data_9top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 1),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 2),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 3),
                          Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_9top10,
           seg = 3,  # Number of axis segments
           title = "9: The Automaticity of Affect for Political Leaders, Groups,
and Issues: An Experimental Test of the Hot Cognition Hypothesis",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 10 of the top 10 ----
data_10top10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is maximum,
the second number is minimum
                          Exclusion = c(3, 0, 3),
                          Sample = c(3, 0, 3),
                          Openness = c(3, 0, 3),
                          Covariates = c(3, 0, 0),
                          Assumptions = c(3,0, 3),
                          Effectsizes = c(3, 0, 0),
                          Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_10top10,
           seg = 3,  # Number of axis segments
           title = "10: Affective and Physiological Responses to the Suffering of
Others: Compassion and Vagal Activity",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 1-10 of the top 10 ----
data_top10radar <- data.frame(Hypotheses = c(3, 0, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2), #
The first number is maximum, the second number is minimum
                              Exclusion = c(3, 0, 1, 0, 0, 3, 1, 1, 1, 0, 1, 3),
                              Sample = c(3, 0, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3),
                              Openness = c(3, 0, 3, 2, 2, 2, 2, 2, 2, 3, 2, 3),
                              Covariates = c(3, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0),
                              Assumptions = c(3, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3),
                              Effectsizes = c(3, 0, 3, 1, 3, 2, 3, 3, 2, 0, 3, 0),
                              Pvalues = c(3, 0, 0, 0, 3, 0, 0, 2, 0, 3, 0, 0),
                              row.names = c("max", "min",
                                            "1: Testosterone and Chess Competition",
                                            "2: Generality of the Automatic Attitude
Activation Effect",
                                            "3: The Scrooge Effect: Evidence That
Mortality Salience Increases Prosocial Attitudes and Behavior",
                                            "4: Varieties of Disgust Faces and the
Structure of Disgust",
                                            "5: Explaining the Enigmatic Anchoring
Effect: Mechanisms of Selective Accessibility",
```

```
                                          "6: Attention in Delay of
Gratification",
                                          "7: Is Empathy-Induced Helping Due to
Self-Other Merging?",
                                          "8: When Approach Motivation and
Behavioral Inhibition Collide: Behavior Regulation Through Stimulus Devaluation",
                                          "9: The Automaticity of Affect for
Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition
Hypothesis",
                                          "10: Affective and Physiological
Responses to the Suffering of Others: Compassion and Vagal Activity")
)

plot.new()
# Define fill colors
colors_fill <- c(scales::alpha("gray", 0.1),
                 scales::alpha("gold", 0.1),
                 scales::alpha("tomato", 0.1),
                 scales::alpha("skyblue", 0.1),
                 scales::alpha("green", 0.1),
                 scales::alpha("pink", 0.1),
                 scales::alpha("purple", 0.1),
                 scales::alpha("orange", 0.1),
                 scales::alpha("black", 0.1),
                 scales::alpha("brown", 0.1))
# Define line colors
colors_line <- c(scales::alpha("darkgray", 0.9),
                 scales::alpha("gold", 0.9),
                 scales::alpha("tomato", 0.9),
                 scales::alpha("royalblue", 0.9),
                 scales::alpha("green", 0.9),
                 scales::alpha("pink", 0.9),
                 scales::alpha("purple", 0.9),
                 scales::alpha("orange", 0.9),
                 scales::alpha("black", 0.9),
                 scales::alpha("brown", 0.9))
# Create plot
radarchart(data_top10radar,
           seg = 3,  # Number of axis segments
           title = "Top 10 - Radar Chart",
           pcol = colors_line,
           pfcol = colors_fill,
           plwd = 4)

# Add a legend
legend(x = 1.5,
       y = 1.2,
       legend = rownames(data_top10radar[-c(1,2),]),
       bty = "n", pch=20 , col = colors_line, cex = 0.9, pt.cex = 2)

# For paper 1 of the bottom 10 ----
data_1bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 1),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 3),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_1bottom10,
           seg = 3,  # Number of axis segments
           title = "1: The Persuasive Effects of a Real and Complex Communication",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)
```

```
# For paper 2 of the bottom 10 ----
data_2bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 1),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 3),
                             Assumptions = c(3,0, 3),
                             Effectsizes = c(3, 0, 0),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_2bottom10,
           seg = 3,  # Number of axis segments
           title = "2: Does the Medium still Matter? The Influence of Gender and
Political Connectedness on Contacting U.S. Public Officials Online and Offline",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 3 of the bottom 10 ----
data_3bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 3),
                             Covariates = c(3, 0, 3),
                             Assumptions = c(3,0, 1),
                             Effectsizes = c(3, 0, 0),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_3bottom10,
           seg = 3,  # Number of axis segments
           title = "3: Confidant Network and Quality of Life of Individuals Aged
50+: The Positive Role of Internet Use",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 4 of the bottom 10 ----
data_4bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 0),
                             Openness = c(3, 0, 1),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 3),
                             Effectsizes = c(3, 0, 0),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_4bottom10,
           seg = 3,  # Number of axis segments
           title = "4: Social Norms and Egalitarian Values Mitigate Authoritarian
Intolerance Toward Sexual Minorities",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 5 of the bottom 10 ----
data_5bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 2),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 0),
```

```
                              Assumptions = c(3,0, 3),
                              Effectsizes = c(3, 0, 2),
                              Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_5bottom10,
           seg = 3,  # Number of axis segments
           title = "5: Personality profiles in substance use disorders: Do they
differ in clinical symptomatology, personality disorders and coping?",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 6 of the bottom 10 ----
data_6bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                              Exclusion = c(3, 0, 0),
                              Sample = c(3, 0, 3),
                              Openness = c(3, 0, 2),
                              Covariates = c(3, 0, 0),
                              Assumptions = c(3,0, 3),
                              Effectsizes = c(3, 0, 2),
                              Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_6bottom10,
           seg = 3,  # Number of axis segments
           title = "6: Pretrial Predictors of Judgments in the O. J. Simpson Case",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 7 of the bottom 10 ----
data_7bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                              Exclusion = c(3, 0, 0),
                              Sample = c(3, 0, 3),
                              Openness = c(3, 0, 2),
                              Covariates = c(3, 0, 0),
                              Assumptions = c(3,0, 3),
                              Effectsizes = c(3, 0, 0),
                              Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_7bottom10,
           seg = 3,  # Number of axis segments
           title = "7: Cultural Factors, Depressive and Somatic Symptoms Among
Chinese American and European American College Students",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 8 of the bottom 10 ----
data_8bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                              Exclusion = c(3, 0, 2),
                              Sample = c(3, 0, 3),
                              Openness = c(3, 0, 2),
                              Covariates = c(3, 0, 0),
                              Assumptions = c(3,0, 3),
                              Effectsizes = c(3, 0, 2),
                              Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_8bottom10,
           seg = 3,  # Number of axis segments
```

```
           title = "8: An emic-etic approach to personality assessment in
predicting social adaptation, risky social behaviors, status striving and social
affirmation",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 9 of the bottom 10 ----
data_9bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 3),
                             Covariates = c(3, 0, 2),
                             Assumptions = c(3,0, 1),
                             Effectsizes = c(3, 0, 3),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_9bottom10,
           seg = 3,  # Number of axis segments
           title = "9: Brazilian Adolescentsâ€™ Just World Beliefs and Its
Relationships with School Fairness, Student Conduct, and Legal Authorities",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 10 of the bottom 10 ----
data_10bottom10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 3),
                             Covariates = c(3, 0, 3),
                             Assumptions = c(3,0, 2),
                             Effectsizes = c(3, 0, 3),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_10bottom10,
           seg = 3,  # Number of axis segments
           title = "10: Cross-level relationships between justice climate and
organizational citizenship behavior: Perceived organizational support as mediator",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 1-10 of the bottom 10 ----
data_bottom10radar <- data.frame(Hypotheses = c(3, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2), # The first number is maximum, the second number is minimum
                              Exclusion = c(3, 0, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0),
                              Sample = c(3, 0, 1, 3, 3, 0, 3, 3, 3, 3, 3, 3),
                              Openness = c(3, 0, 2, 2, 3, 1, 2, 2, 2, 2, 3, 3),
                              Covariates = c(3, 0, 0, 3, 3, 0, 0, 0, 0, 0, 2, 3),
                              Assumptions = c(3, 0, 3, 3, 1, 3, 3, 3, 3, 3, 1, 2),
                              Effectsizes = c(3, 0, 2, 0, 0, 0, 2, 2, 0, 2, 3, 3),
                              Pvalues = c(3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
                              row.names = c("max", "min",
                                            "1: The Persuasive Effects of a Real
and Complex Communication",
                                            "2: Does the Medium still Matter? The
Influence of Gender and Political Connectedness on Contacting U.S. Public Officials
Online and Offline",
                                            "3: Confidant Network and Quality of
Life of Individuals Aged 50+: The Positive Role of Internet Use",
                                            "4: Social Norms and Egalitarian Values
Mitigate Authoritarian Intolerance Toward Sexual Minorities",
```

```
                                                "5: Personality profiles in substance
use disorders: Do they differ in clinical symptomatology, personality disorders and
coping?",
                                                "6: Pretrial Predictors of Judgments in
the O. J. Simpson Case",
                                                "7: Cultural Factors, Depressive and
Somatic Symptoms Among Chinese American and European American College Students",
                                                "8: An emic-etic approach to
personality assessment in predicting social adaptation, risky social behaviors,
status striving and social affirmation",
                                                "9: Brazilian Adolescentsâ€™ Just World
Beliefs and Its Relationships with School Fairness, Student Conduct, and Legal
Authorities",
                                                "10: Cross-level relationships between
justice climate and organizational citizenship behavior: Perceived organizational
support as mediator")
)

plot.new()
# Define fill colors
colors_fill <- c(scales::alpha("gray", 0.1),
                 scales::alpha("gold", 0.1),
                 scales::alpha("tomato", 0.1),
                 scales::alpha("skyblue", 0.1),
                 scales::alpha("green", 0.1),
                 scales::alpha("pink", 0.1),
                 scales::alpha("purple", 0.1),
                 scales::alpha("orange", 0.1),
                 scales::alpha("black", 0.1),
                 scales::alpha("brown", 0.1))
# Define line colors
colors_line <- c(scales::alpha("darkgray", 0.9),
                 scales::alpha("gold", 0.9),
                 scales::alpha("tomato", 0.9),
                 scales::alpha("royalblue", 0.9),
                 scales::alpha("green", 0.9),
                 scales::alpha("pink", 0.9),
                 scales::alpha("purple", 0.9),
                 scales::alpha("orange", 0.9),
                 scales::alpha("black", 0.9),
                 scales::alpha("brown", 0.9))
# Create plot
radarchart(data_bottom10radar,
           seg = 3,  # Number of axis segments
           title = "Bottom 10 - Radar Chart",
           pcol = colors_line,
           pfcol = colors_fill,
           plwd = 4)

# Add a legend
legend(x = 1.5,
       y = 1.2,
       legend = rownames(data_bottom10radar[-c(1,2),]),
       bty = "n", pch=20 , col = colors_line, cex = 0.9, pt.cex = 2)

# For paper 1 of the center 10 ----
data_1center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 3),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 1),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 0)
)

plot.new()
```

```
radarchart(data_1center10,
           seg = 3,  # Number of axis segments
           title = "1: A research experience for American Indian undergraduates:
Utilizing an actorâ€“partner interdependence model to examine the studentâ€“mentor
dyad",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 2 of the center 10 ----
data_2center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 1),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 1),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_2center10,
           seg = 3,  # Number of axis segments
           title = "2: Environmental Resources and the Posttreatment Functioning of
Alcoholic Patients",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 3 of the center 10 ----
data_3center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 3),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 1),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_3center10,
           seg = 3,  # Number of axis segments
           title = "3: Self-Reported Attachment Patterns and Rorschach-Related
Scores of Ego Boundary, Defensive Processes, and Thinking Disorders",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 4 of the center 10 ----
data_4center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 1),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 3),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 2)
)

plot.new()
radarchart(data_4center10,
           seg = 3,  # Number of axis segments
           title = "4: A license to speak up: Outgroup minorities and opinion
expression",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)
```

```
# For paper 5 of the center 10 ----
data_5center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 1),
                             Effectsizes = c(3, 0, 3),
                             Pvalues = c(3, 0, 2)
)

plot.new()
radarchart(data_5center10,
           seg = 3,  # Number of axis segments
           title = "5: When Stigma Confronts Stigma: Some Conditions Enhancing a
Victimâ€™s Tolerance of Other Victims",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 6 of the center 10 ----
data_6center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 2),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 0),
                             Assumptions = c(3,0, 3),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_6center10,
           seg = 3,  # Number of axis segments
           title = "6: Consideration of future consequences scale: Confirmatory
Factor Analysis",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 7 of the center 10 ----
data_7center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 0),
                             Sample = c(3, 0, 3),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 3),
                             Assumptions = c(3,0, 3),
                             Effectsizes = c(3, 0, 2),
                             Pvalues = c(3, 0, 2)
)

plot.new()
radarchart(data_7center10,
           seg = 3,  # Number of axis segments
           title = "7: Social Identity, Modern Sexism, and Perceptions of Personal
and Group Discrimination by Women and Men",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 8 of the center 10 ----
data_8center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                             Exclusion = c(3, 0, 3),
                             Sample = c(3, 0, 2),
                             Openness = c(3, 0, 2),
                             Covariates = c(3, 0, 3),
                             Assumptions = c(3,0, 1),
```

```
                              Effectsizes = c(3, 0, 0),
                              Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_8center10,
           seg = 3,  # Number of axis segments
           title = "8: Acute Thoughts, Exercise Consistency, and Coping Self-
Efficacy",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 9 of the center 10 ----
data_9center10 <- data.frame(Hypotheses = c(3, 0, 2), # The first number is
maximum, the second number is minimum
                              Exclusion = c(3, 0, 1),
                              Sample = c(3, 0, 3),
                              Openness = c(3, 0, 2),
                              Covariates = c(3, 0, 0),
                              Assumptions = c(3,0, 3),
                              Effectsizes = c(3, 0, 2),
                              Pvalues = c(3, 0, 0)
)

plot.new()
radarchart(data_9center10,
           seg = 3,  # Number of axis segments
           title = "9: Tyramine, a new clue to disinhibition and sensation
seeking?",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 10 of the center 10 ----
data_10center10 <- data.frame(Hypotheses = c(3, 0, 3), # The first number is
maximum, the second number is minimum
                              Exclusion = c(3, 0, 0),
                              Sample = c(3, 0, 3),
                              Openness = c(3, 0, 3),
                              Covariates = c(3, 0, 3),
                              Assumptions = c(3,0, 3),
                              Effectsizes = c(3, 0, 3),
                              Pvalues = c(3, 0, 2)
)

plot.new()
radarchart(data_10center10,
           seg = 3,  # Number of axis segments
           title = "10: The Effects of Race, Weight, and Gender on Evaluations of
Writing Competence",
           pfcol = scales::alpha("gray", 0.3),
           plwd = 2)

# For paper 1-10 of the center 10 ----
data_center10radar <- data.frame(Hypotheses = c(3, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2,
3), # The first number is maximum, the second number is minimum
                                 Exclusion = c(3, 0, 0, 1, 0, 1, 0, 0, 0, 3, 1, 0),
                                 Sample = c(3, 0, 3, 3, 3, 3, 3, 2, 3, 2, 3, 3),
                                 Openness = c(3, 0, 3, 2, 3, 2, 2, 2, 2, 2, 2, 3),
                                 Covariates = c(3, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0,
3),
                                 Assumptions = c(3, 0, 1, 1, 1, 3, 1, 3, 3, 1,
3,3),
                                 Effectsizes = c(3, 0, 2, 2, 2, 2, 3, 2, 2, 0, 2,
3),
                                 Pvalues = c(3, 0, 0, 0, 0, 2, 2, 0, 2, 0, 0, 2),
                                 row.names = c("max", "min",
```

```
                                                "1: A research experience for
American Indian undergraduates: Utilizing an actorâ€“partner interdependence model
to examine the studentâ€“mentor dyad",
                                                "2: Environmental Resources and the
Posttreatment Functioning of Alcoholic Patients",
                                                "3: Self-Reported Attachment
Patterns and Rorschach-Related Scores of Ego Boundary, Defensive Processes, and
Thinking Disorders",
                                                "4: A license to speak up: Outgroup
minorities and opinion expression",
                                                "5: When Stigma Confronts Stigma:
Some Conditions Enhancing a Victimâ€™s Tolerance of Other Victims",
                                                "6: Consideration of future
consequences scale: Confirmatory Factor Analysis",
                                                "7: Social Identity, Modern Sexism,
and Perceptions of Personal and Group Discrimination by Women and Men",
                                                "8: Acute Thoughts, Exercise
Consistency, and Coping Self-Efficacy",
                                                "9: Tyramine, a new clue to
disinhibition and sensation seeking?",
                                                "10: The Effects of Race, Weight,
and Gender on Evaluations of Writing Competence")
)

plot.new()
# Define fill colors
colors_fill <- c(scales::alpha("gray", 0.1),
                 scales::alpha("gold", 0.1),
                 scales::alpha("tomato", 0.1),
                 scales::alpha("skyblue", 0.1),
                 scales::alpha("green", 0.1),
                 scales::alpha("pink", 0.1),
                 scales::alpha("purple", 0.1),
                 scales::alpha("orange", 0.1),
                 scales::alpha("black", 0.1),
                 scales::alpha("brown", 0.1))
# Define line colors
colors_line <- c(scales::alpha("darkgray", 0.9),
                 scales::alpha("gold", 0.9),
                 scales::alpha("tomato", 0.9),
                 scales::alpha("royalblue", 0.9),
                 scales::alpha("green", 0.9),
                 scales::alpha("pink", 0.9),
                 scales::alpha("purple", 0.9),
                 scales::alpha("orange", 0.9),
                 scales::alpha("black", 0.9),
                 scales::alpha("brown", 0.9))
# Create plot
radarchart(data_center10radar,
           seg = 3,  # Number of axis segments
           title = "Center 10 - Radar Chart",
           pcol = colors_line,
           pfcol = colors_fill,
           plwd = 4)

# Add a legend
legend(x = 1.5,
       y = 1.2,
       legend = rownames(data_center10radar[-c(1,2),]),
       bty = "n", pch=20 , col = colors_line, cex = 0.9, pt.cex = 2)
```