Meta-Psychology, 2019, vol 3, MP.2018.878,
https://doi.org/10.17605/MP.2018.878
Article type: Original Article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Rickard Carlsson
Reviewed by: Joseph Hilgard, Moritz Heene, Jens Agerström
Analysis reproduced by: Rickard Carlsson
All supplementary files can be accessed at the OSF project page:
https://doi.org/10.17605/OSF.IO/H2RX7

Has the evidence for moral licensing been inflated by publication bias?

Niclas Kuper
University of Hamburg, Germany

Antonia Bott
University of Hamburg, Germany

Moral licensing describes the phenomenon that displaying moral behavior can lead to subsequent immoral behavior. This is usually explained by the idea that an initial moral act affirms the moral self-image and hence licenses subsequent immoral acts. Previous meta-analyses on moral licensing indicate significant overall effects of $d > .30$. However, several large replication studies have either not found the effect or reported a substantially smaller effect size. The present article investigated whether this can be attributed to publication bias. Datasets from two previous meta-analyses on moral licensing were compared and when necessary modified. The larger dataset was used for the present analyses. Using PET-PEESE and a three-parameter-selection-model (3-PSM), we found some evidence for publication bias. The adjusted effect sizes were reduced to $d = -0.05$, $p = .64$ and $d = 0.18$, $p = .002$, respectively. While the first estimate could be an underestimation, we also found indications that the second estimate might exaggerate the true effect size. It is concluded that both the evidence for and the size of moral licensing effects has likely been inflated by publication bias. Furthermore, our findings indicate that culture moderates the moral licensing effect. Recommendations for future meta-analytic and empirical work are given. Subsequent studies on moral licensing should be adequately powered and ideally pre-registered.

Keywords: moral licensing, meta-analysis, replication crisis, publication bias

Moral licensing theory postulates that initially displaying a moral action increases the probability of subsequent behavior that is immoral, unethical or otherwise problematic (Merritt, Effron, & Monin, 2010). This phenomenon has been investigated in several studies across various life domains. For instance, previous gender-egalitarian acts were demonstrated to result in an increased likelihood of subsequent gender-discriminatory behavior in form of characterizing stereotypically masculine jobs as better suited for men than women (Monin & Miller, 2001). Similarly, participants who were initially given the chance to purchase green products subsequently shared less money in an interpersonal interaction task and stole more money in a self-gratification paradigm (Mazar & Zhong, 2010). Strikingly, the moral licensing effect appears to be at odds with several prominent psychological findings and theories which imply a human striving for consistency (Blanken, van de Ven, Zeelenberg, & Meijers, 2014). For instance, self-perception theory (Bem, 1972) posits

Correspondence concerning this article should be addressed to Niclas Kuper via niclas.kuper@gmail.com

that people constantly observe their own behavior in order to draw conclusions about their own attitudes. These inferred attitudes strongly influence subsequent behavior and are thought to establish consistency. The observed *inconsistency* in human behavior described as moral licensing, by contrast, was explained by drawing upon a moral self-regulation framework (Sachdeva, Iliev, & Medin, 2009). Specifically, the predicted costs associated with a future ethical deed were argued to be counted against the gains in moral self-concept acquired by previous moral actions. Thus, whenever faced with the uncertainty whether their next behavior might appear morally questionable, an individual's clean moral record might license an ensuing immoral behavior - without threatening the previously verified ethical self-concept (Monin & Miller, 2001; Sachdeva et al., 2009). For illustration, a personnel manager who has just objected a set of racist statements might feel licensed to prefer a White over an equally qualified Black applicant without having to readapt his non-racist self-concept. Similarly, an initial moral transgression was reasoned to threaten an individual's perceived moral self-concept (Zhong & Liljenquist, 2006). In order to reinstate an internal balance and to regain confidence in the moral self-concept, the individual might therefore engage in an ensuing compensatory moral action (Sachdeva et al., 2009). For instance, writing a story about oneself using negative words (such as greedy, disloyal, selfish) as compared to neutral words (book, keys, house) increased the mean amount of money participants indicated they would donate to charity (Sachdeva et al., 2009). This set of compensatory behaviors following a moral transgression is known as *moral cleansing* (Sachdeva et al., 2009). Thus, both the moral cleansing and the moral licensing effect could be argued to reflect the outcome of constant moral cost-benefit calculations.

Moderators of the moral licensing effect are largely unknown (Blanken, van de Ven, & Zeelenberg, 2015), but Simbrunner and Schlegelmilch (2017) recently provided evidence for two potential moderators: (1) The culture of the studied population and (2) the type of control condition (neutral vs. immoral previous behavior). Specifically, they argued that the cultural background of participants shapes their moral standards by processes of socialization and thus influences their moral self-concept. Culture should hence moderate the moral licensing effect (Simbrunner & Schlegelmilch, 2017). Additionally, they showed that the effect was larger when a previous moral behavior was contrasted with a previous immoral behavior instead of a neutral control condition.

Replication issues

Two recent meta-analyses have been conducted on the moral licensing effect. Blanken et al. (2015) found an effect size of $d = 0.31$ over 91 studies and more recently, Simbrunner and Schlegelmilch (2017) reported an overall effect size of $d = 0.32$ over 106 studies. Despite these seemingly encouraging findings, there have been concerns about the replicability of moral licensing effects. Specifically, two highly powered sets of studies ($n = 801$ and $n = 1{,}274$) found no evidence for a moral licensing effect (Blanken et al., 2014; Urban, Bahník, & Kohlová, 2017). An even larger replication attempt was conducted as part of a many-labs replication project. The study found the expected licensing effect with $n = 3{,}134$, but the effect size was less than half of the estimates derived from recent meta-analyses ($d = 0.15$; Ebersole et al., 2016).

How can these replication issues and reduced effect sizes in large studies be reconciled with the results of the meta-analyses? This question touches upon the larger issue of a “crisis of reproducibility” in psychology (Pashler & Wagenmakers, 2012). The replicability of psychological studies has recently been empirically estimated to be only around 40%, although this estimate varies across subfields of psychology (Open Science Collaboration, 2015). In addition to this, questions were raised about the replicability of several well-established effects that were studied in hundreds of experiments. A well-known example of this is the ego depletion literature (Baumeister, Bratslavsky, Muraven, & Tice, 1998), where a meta-analysis over almost 200 studies reported an effect size of $d = 0.62$ (Hagger, Wood, Stiff, & Chatzisarantis, 2010). However, a re-analysis indicated that the effect is not as robust as previously thought (Carter & McCullough, 2014; see below) and two recent highly powered replication projects found no evidence for an effect (Hagger et al., 2016; Etherton et al., 2018).

At least two core reasons have been identified for the surprising lack of reproducibility in some areas of psychology: (1) publication bias - the preferential publication of significant results – and (2) questionable research practices (QRPs) that inflate the type-1-error rate. Publication bias (Rothstein, Sutton, & Borenstein, 2005) has been a well-known problem in psychology for several decades (e.g., Greenwald, 1975) and the replication crisis has raised awareness for the issue (e.g., Ferguson & Heene, 2012; Francis, 2012; Franco, Malhotra, & Simonovits, 2014; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014). It represents a major threat to conventional meta-analyses given that publication bias which is unaccounted for can lead to type-1-error rates close to 100% and substantial effect size estimates in the absence of an effect (e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2018). In addition, researcher degrees of freedom allow for QRPs that artificially decrease *p*-values, with the explicit or implicit goal to obtain $p < .05$. Examples of such QRPs include optional stopping, dropping experimental conditions, changing the dependent variable and selective exclusion of outliers (Simmons, Nelson, & Simonsohn, 2011).

Testing and correcting for publication bias

In order to test whether publication bias can explain recent replication issues of moral licensing, we attempt to both detect and correct for publication bias in the moral licensing literature. Blanken et al. (2015) have already provided evidence for publication bias by regressing effect sizes on their standard errors. They found that larger standard errors were associated with larger effect sizes, which suggests the presence of publication bias (Egger, Smith, Schneider, & Minder, 1997; see below). However, they restricted this analysis to published studies only instead of considering the entire dataset. This approach might miss publication bias that persisted even after the inclusion of some unpublished studies. In addition, they have graphically used the trim and fill method (Duval & Tweedie, 2000) to correct for publication bias, which has been criticized (Moreno et al., 2009). Finally, they showed that the mean effect size of published studies is larger than that of unpublished studies, again suggesting the presence of publication bias (Blanken et al., 2015). Contrary to this, Simbrunner and Schlegelmilch (2017) showed that the effect of publication status disappeared in a moderator-analysis which included other predictors of moral licensing. Using Rosenthal's (1979) fail-safe N, they demonstrate that 4,531 nonsignificant unpublished studies would be required to reduce the overall result to nonsignificance. Simbrunner and Schlegelmilch (2017) conclude that their result is robust to publication bias.

We seek to expand on these initial statistical tests for publication bias in the moral licensing literature by testing and correcting for publication bias in the entire dataset including unpublished studies. Since not all unpublished studies can be obtained in meta-analyses, datasets including unpublished work are still likely to be influenced by publication bias. We implemented methods to correct for publication bias which have shown to perform best in simulation studies. The previously employed methods trim and fill and fail-safe N have both been shown to possess less than optimal properties. Specifically, trim and fill substantially overestimates the true effect size in the case of publication bias (Moreno et al., 2009; Carter et al., 2018). Fail-safe N, on the other hand, suffers from a fundamental error: It assumes that the mean Z-score of unpublished studies is zero, while it would actually be negative in the case of an overall effect size of zero paired with publication bias (Scargle, 2000; Schonemann, & Scargle, 2008). Furthermore, even improved fail-safe N calculations that attempt to circumvent this flaw have at least two additional issues: (1) QRPs that increase the type-1-error rate artificially increase the estimated fail-safe N (Simonsohn, Nelson, & Simmons, 2014) and (2) fail-safe N is merely concerned with statistical significance and not effect sizes. Meta-analyses, however, are usually concerned with effect size estimation and not just the test of (nil-)null hypotheses. The use of fail-safe N has thus been discouraged in favor of other available methods (Becker, 2005). We will focus on two such methods: PET-PEESE and the three-parameter selection model (3-PSM).

PET-PEESE

PET-PEESE (Stanley & Doucouliagos, 2014) can be used both as a test for publication bias and as a

way to correct for publication bias. It models the relationship between effect sizes and their standard errors using a weighted linear regression. A significant slope indicates the existence of publication bias given that in most cases no such relationship should exist without publication bias. The selective omission of small studies with small effect sizes, however, leads to a positive relationship between effect size and standard error. The intercept of the weighted regression provides an estimate of the corrected mean effect size. If the intercept is significant, the model is re-estimated with the variance instead of the standard error as the predictor given that this model is less biased in the presence of a non-zero effect. PET-PEESE has been famously applied to the ego-depletion (Baumeister et al., 1998) literature. Whereas a meta-analysis on ego-depletion identified a mean effect size of d = 0.62 (Hagger et al., 2010), PET-PEESE suggested an effect that is not different from zero (Carter & McCullough, 2014). In line with the results from PET-PEESE, subsequent highly-powered replications of ego depletion indeed did not find an effect (Hagger et al., 2016; Etherton et al., 2018). PET-PEESE outperforms traditional methods to correct for publication bias such as trim and fill (see simulations from Carter et al., 2018), although it can have both upward and downward biases in certain situations. For instance, it can overestimate effect sizes of zero in the case of extreme heterogeneity together with publication bias. This is worsened when only few studies are included in the meta-analysis (Carter et al., 2018; Stanley, 2017). Nevertheless, when heterogeneity and publication bias are present, it outperforms other frequently used methods including trim and fill and p-curve in terms of both effect size estimation and type-1-error control (see simulations from Carter et al., 2018).

3-PSM

Selection models were initially proposed by Hedges (1984) and extended by Iyengar and Greenhouse (1988) as well as Vevea and Hedges (1995). The model from Iyengar and Greenhouse (1988) simultaneously estimates (1) a mean effect size, (2) between-study heterogeneity and (3) a probability that nonsignificant studies are published. These three parameters are estimated by optimizing the joint likelihood function. A likelihood ratio test is employed to investigate whether modelling publication bias increases model fit. The model was further extended by Vevea and Hedges (1995) to incorporate moderator analyses into the selection model. The 3-PSM performed better than any other method in the simulations from Carter et al. (2018). Its mean effect size and type-1-error rate were nominal or close to nominal in most cases. However, it should be noted, that in the simulations the data-generating mechanism exactly matched the model of the 3-PSM. In cases when its assumptions, such as the normal distribution of heterogeneous effect sizes, are violated, the performance might be worse (see Hedges & Vevea, 1996). Hence, we decided to implement both the 3-PSM and PET-PEESE and compare their results.

In sum, previous work has indicated the presence of publication bias in the moral licensing literature. However, this was in part restricted to a subset of the data and in part limited by the use of suboptimal methods. We want to add to the literature by analyzing the most recent Simbrunner and Schlegelmilch (2017) dataset with state-of-the-art methods to detect and correct for publication bias. To this end, we (1) compare the Blanken et al. (2015) and the Simbrunner and Schlegelmilch (2017) datasets to confirm that the same experiments were coded identically and to correct errors if necessary, (2) replicate the Simbrunner and Schlegelmilch (2017) results and (3) test and correct for publication bias using PET-PEESE and the 3-PSM.

Comparison between the two meta-analyses

A comparison between the two meta-analyses indicated that 93% of the effect sizes (k = 85) associated with the same studies were identical. We identified 6 effect sizes that were coded differently in the two datasets. For discrepancies and their resolution see Table 1. The largest discrepancy emerged in the case of a study (Mazar & Zhong, 2010) where an effect size of d = 0.52 was coded as d = 3.2 and d = 3.5 by Simbrunner and Schlegelmilch (2017). This error could be attributed to the fact that the authors of this study erroneously reported SE as SD (see Mazar & Zhong, 2010). The smaller effect size is consistent with the

reported t-values. Regarding the sample sizes, 93% were coded to be identical. For 95% of studies, the difference in coded sample size was smaller than 2. Six studies for which the sample sizes differed emerged. For discrepancies and their resolutions see Table 1.

In addition, we examined effects that were included in one meta-analysis but not in the other. In most of these cases, a report was indeed included in only one meta-analysis. In some cases, however, the report was included in both meta-analyses, but additional effects were coded in one of them. These additional effect sizes were removed because they either were duplicates or violated the independence assumption. Furthermore, we excluded three studies (7 effect sizes) from the Simbrunner and Schlegelmilch (2017) dataset since they did not assess moral licensing in the traditional sense (see Table 1).

Ensuring the independence of effect sizes

In the modified datasets, 100% of the effect sizes and 100% of the sample sizes were identical. Our analyses focus on the larger Simbrunner and Schlegelmilch (2017) dataset. However, we had to make further changes to this dataset to ensure that the independence assumption of traditional meta-analyses is satisfied. Specifically, some studies within the dataset reported multiple outcome variables from the same experiment which therefore had to be aggregated. Furthermore, both the Blanken et al. (2015) and the Simbrunner and Schlegelmilch (2017) meta-analyses coded two separate effects when two experimental conditions were contrasted with the same control condition, likewise violating effect size independence. For details and our courses of action, see Table 2. After independence was ensured, $k = 76$ effect sizes remained.

Replication of Simbrunner and Schlegelmilch (2017)

First, we attempted to replicate the essential findings reported in the meta-analysis. Analyses were implemented using the metafor-package (Viechtbauer, 2010) of the open source statistics software R (R Core Team, 2016). A random-effects meta-analysis indicated a mean effect size of $d = 0.27$, [0.19; 0.35], $Z = 6.57$, $p < .001$, which is slightly smaller than the $d = 0.32$ reported by Simbrunner and Schlegelmilch (2017).

Table 1.
Discrepancies in effect size/sample size/moderator coding and study inclusion between the meta-analyses of Blanken et al. (2015) and Simbrunner and Schlegelmilch (2017) and their resolutions.

Discrepancy	Study	Action taken
ES coding & Inclusion	Blanken et al. (2014), study 3: Coded as two studies by Simbrunner and Schlegelmilch (2017). Two dependent variables were reported. An average of the two ES is similar to the ES reported in Blanken et al. (2015). This average was included.	Given that these ES are not independent, we decided to average them and code them as one. Sign reversed (see below).
ES coding	Blanken et al. (2014), study 1-3: ES were reported to be in the opposite direction of that reported by Blanken et al. (2015) and Simbrunner and Schlegelmilch (2017).	The direction of the ES was reversed in both datasets, consistent with the original Blanken et al. (2014) report.
ES coding & Inclusion	Effron, Monin, & Miller (2012), study 1: Simbrunner and Schlegelmilch (2017) included study twice with different ES.	Instead of including both dependent ES separately, we averaged them, after which the ES was identical to that of Blanken et al. (2015).

Discrepancy	Study	Action taken
ES coding	Jordan, Mullen, & Murnighan (2011), study 2: Two control conditions (neutral vs. immoral) available, neutral coded by Blanken et al. (2015), immoral by Simbrunner & Schlegelmilch (2017).	ES from neutral control group imputed in the Simbrunner and Schlegelmilch (2017) dataset.
ES coding	Jordan et al. (2011), study 3: ES difference of .03 between the two meta-analyses. Negligible given that d = 1.00 in this case.	The ES was recalculated, and both ES were coded as d = .994.
ES coding	Mazar and Zhong (2010), study 3: recalculated d = .52, coded as d = 3.2 by Simbrunner and Schlegelmilch (2017). Coded correctly as d = .53 in Blanken et al. (2015).	ES corrected in the Simbrunner and Schlegelmilch (2017) dataset to be d = .53.
ES coding	Young, Chakroff, and Tom (2012), study 1: d = -.41 coded correctly by Simbrunner and Schlegelmilch (2017), coded as d = .41 by Blanken et al. (2015).	ES corrected in the Blanken et al. (2015) dataset.
N coding	Blanken, van de Ven, and Zeelenberg (2012), study 6: n = 64 - first reported in Blanken et al. (2015), coded as n = 54 in Simbrunner and Schlegelmilch (2017).	N was replaced in Simbrunner and Schlegelmilch (2017) and the SE was recalculated.
N coding	Blanken et al. (2014), study 4: Simbrunner and Schlegelmilch (2017) coded n = 614, Blanken et al. (2015) coded n = 567. The original report (Blanken et al. (2014)) does not clearly report n = 567.	Despite n being closer to 614 in the original report, we imputed n = 567 from the Blanken et al. (2015) meta-analysis.
N coding	Jordan et al. (2011), study 2 & study 3: n = 68 and n = 76, respectively, coded correctly by Blanken et al. (2015) and coded incorrectly as n = 84 and n = 84 by Simbrunner and Schlegelmilch (2017).	N was replaced in the Simbrunner and Schlegelmilch (2017) dataset and the SE was recalculated.
N coding	Mazar and Zhong (2010), study 3: n = 81 coded as n = 80 by Simbrunner and Schlegelmilch (2017).	N was replaced, and SE was recalculated.
N coding	Monin and Miller (2001), study 3: n = 21 coded as n = 20 by Simbrunner and Schlegelmilch (2017).	N was replaced, and SE was recalculated.
Inclusion	Jordan et al. (2011), study 1: Study 1 was coded by Simbrunner and Schlegelmilch (2017). However, study 1 only includes moral identity as DV and not behavior. It is hence not a measure of moral licensing.	The study was omitted.
Inclusion	Effron (2014), study 1 and 2: Simbrunner and Schlegelmilch (2017) included these studies although they do not assess moral licensing. The studies dealt with meta-perceptions of moral credentials related to prior actions.	Both ES were omitted.

(continued)

Discrepancy	Study	Action taken
Inclusion	Kouchaki (2011), study 1 to study 4: Simbrunner and Schlegelmilch (2017) included these studies although they do not assess moral licensing in the traditional sense. The studies investigated vicarious licensing, i.e., credentials established through others' actions.	All ES were omitted.
Inclusion	Mazar and Zhong (2010), study 3: Study 3 was coded twice by Simbrunner and Schlegelmilch (2017).	The ES was recalculated (see above) and the second ES was omitted.
Inclusion	Monin and Miller (2001), study 3: Study 3 was coded twice by Blanken et al. (2015).	The second ES was omitted.
Moderator coding	Jordan et al. (2011), study 3: Immoral control condition coded as immoral by Simbrunner & Schlegelmilch (2017) but as neutral by Blanken et al. (2015).	Moderator coding corrected in the Blanken et al. (2015) dataset.
Moderator coding	Mazar and Zhong (2010), study 1 & 2: Neutral control condition coded as neutral by Blanken et al. (2015) but as immoral by Simbrunner & Schlegelmilch (2017).	Moderator coding corrected in the Simbrunner & Schlegelmilch (2017) dataset.

Note. Inclusion: The inclusion of the study led to a discrepancy between the meta-analyses. ES coding & N coding: The effect sizes or the sample sizes were coded inconsistently. Moderator coding: The moderator "type of control condition" was coded inconsistently.

There was evidence for substantial heterogeneity, $I^2 = 0.26$, $Q(75) = 175.77$, $p < .001$. We further replicated the moderator-analysis using culture (Europe vs. North-America vs. South-East Asia) and type of comparison (neutral vs. immoral control condition) as predictors. As reported by Simbrunner and Schlegelmilch (2017), culture had a significant effect, with North-American samples showing significantly larger effect sizes than South-East Asian samples, $\beta = .71$, $Z = 2.34$, $p = .019$. However, the difference between European and South-East Asian samples was no longer significant in our analyses, $\beta = .54$, $Z = 1.76$, $p = .078$. This is attributable to an increased standard error in our corrected dataset. Specifically, Simmbrunner and Schlegelmilch (2017) coded multiple out-comes from the same sample as separate effect sizes while we averaged over outcomes to ensure independence. This left us with only one study from South-East Asia (see Table 2).

Table 2.
Details on the exclusion of non-independent studies.

Non-independent studies	Course of Action
Blanken et al. (2012), study 1 and study 2: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Blanken et al. (2012), study 4 and study 5: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Blanken et al. (2012), study 6 and study 7: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Blanken et al. (2012), study 8 and study 9: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]

Non-independent studies	Course of Action
Blanken et al. (2012), study 10 and study 11: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Blanken et al. (2012), study 12 and study 13: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Blanken et al. (2012), study 14 and study 15: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Blanken et al. (2012), study 16 and study 17: two experimental groups were compared to one control group	Recalculated ES and SE using only the reported ES[1]
Bradley-Geist et al. (2010), study 1 and study 2: two experimental groups were compared to one control group	Recalculated ES and SE using reported means and standard deviations[2]
Bradley-Geist et al. (2010), study 3 and study 4: two experimental groups were compared to one control group	Recalculated ES and SE using reported means and standard deviations[2]
Meijers et al. (2014), study 1 and study 2: two experimental groups were compared to one control group	Recalculated ES and SE using reported means and standard deviations[2]
Meijers et al. (2014), study 3 and study 4: two experimental groups were compared to one control group	Recalculated ES and SE using reported means and standard deviations[2]
Simbrunner and Schlegelmilch (2016), study 2.1 to study 2.5 (n = 57): multiple effect sizes from the same sample (P. Simbrunner, personal communication, August 22, 2018).	Recalculated ES and SE by averaging over effect sizes and their standard errors[3]
Simbrunner and Schlegelmilch (2016), study 3.1 to study 3.5 (n = 111): multiple effect sizes from the same sample (P. Simbrunner, personal communication, August 22, 2018).	Recalculated ES and SE by averaging over effect sizes and their standard errors[3]

Note. [1]Recalculated the effect size and standard error: Calculated a theoretical mean and standard deviation of the pooled experimental group based on the reported effect sizes only. Assuming standard errors of 1 and a mean of 0 in the control group. [2]Recalculated the effect size and standard error by determining the exact mean + standard deviation of the pooled experimental group and comparing it with the control group. [3]Averaged over all effect sizes and their standard errors given that multiple outcomes were reported.

The more highly powered contrast between North American and European samples was significant, β = .17, Z = 2.17, p = .030.

The effect of type of comparison – neutral control conditions leading to smaller licensing effects than immoral control conditions – was significant, β = -.19, Z = -1.97, p = .049. Nevertheless, the corrections reported in Table 1 reduced the moderating influence of type of comparison, which was substantially higher in the uncorrected dataset, β = -.42, Z = -3.72, p < .001. This is largely attributable to the accidental inclusion of two effect sizes with d > 3 in the immoral condition (see Table 1).

Publication bias analyses

Over all 76 effect sizes, PET-PEESE indicated a significant positive slope for the standard error, β = 1.36, t(74) = 2.73, p = .008. The intercept, representing the corrected effect size, was not different from zero, β = -.05, 95% CI [-0.26; 0.16], t(74) = -0.46, p = .64. A visual inspection of the funnel plot supports these findings: There is clear indication of asymmetry with smaller studies yielding large effects and larger studies showing smaller effect sizes (see Figure 1). A 3-PSM (using the weightr R package; Coburn & Vevea, 2017) over all studies indicated a significant corrected effect size of d = 0.18, 95% CI [0.06; 0.29], Z = 3.11, p = .002. There

was no significant improvement in fit when publication bias was modelled vs. when it was not, $\chi^2(1) = 3.42$, $p = .065$. In order to test the robustness of these findings, we repeated the analyses after simultaneously including relevant moderators.

A meta-regression using type of comparison, culture and – resembling PET-PEESE – the standard error as predictors indicated significant effects of culture: North-America vs. South-East-Asia, β = .66, Z = 2.26, $p = .024$, North-America vs. Europe, β = .15, Z = 1.98, $p = .048$, and the standard error, β = 1.20, Z = 2.00, $p = .045$. Type of comparison was no longer a significant predictor, β = -.14, Z = -1.49, $p = .14$.

Table 3.
Publication bias analyses.

Sample	Naïve rma	PET-PEESE	3-PSM
Entire dataset ($k = 76$)	$d = .27$ 95% CI [0.19; 0.35], $Z = 6.57$, $p < .001$	$d = -.05$ 95% CI [-0.26; 0.16], $t(74) = -0.46$, $p = .64$	$d = .18$ 95% CI [0.06 ; 0.29], $Z = 3.11$, $p = .002$
		$b = 1.36$, $t(74) = 2.73$, $p = .008$	fit improvement: $\chi^2(1) = 3.42$, $p = .065$
North America ($k = 40$)	$d = .38$ 95% CI [0.27; 0.48], $Z = 7.01$, $p < .001$	$d = -.13$ 95% CI [-0.37; 0.12], $t(38) = -1.07$, $p = .29$	$d = .31$ 95% CI [0.15 ; 0.47], $Z = 3.76$, $p < .001$
		$b = 2.12$, $t(38) = 3.82$, $p < .001$	fit improvement: $\chi^2(1) = 1.02$, $p = .31$
Europe ($k = 30$)	$d = .21$ 95% CI [0.09 ; 0.32], $Z = 3.45$, $p < .001$	$d = .10$ 95% CI [-0.28; 0.48], $t(28) = 0.54$, $p = .59$	$d = .11$ 95% CI [-0.03; 0.24], $Z = 1.53$, p = .12
		$b = .49$, $t(28) = 0.55$, $p = .58$	fit improvement: $\chi^2(1) = 2.40$, $p = .12$
South-East Asia ($k = 1$)	$d = -.37$ 95% CI [-0.75; 0.004], $Z = -1.94$, $p = .052$	NA	NA

Note. Naïve rma = conventional random effects meta-analysis. PET-PEESE: *d* denotes the intercept/corrected effect size, *b* denotes the slope/the test for potential publication bias. 3-PSM: *d* denotes the corrected effect size. The fit-improvement compares the model fit with a baseline model that does not estimate publication bias. NA: Not applicable.

A 3-PSM including type of comparison and culture as moderators indicated significant effects of world region: North-America vs. South-East Asia, β = .63, Z = 2.31, $p = .021$, North-America vs. Europe, β = .15, Z = 2.14, $p = .033$; with no significant improvement in fit after publication bias is modelled, $\chi^2(1) = 2.77$, $p = .096$. Again, type of comparison was not significant anymore, β = -.17, Z = -1.89, $p = .059$.

We further repeated the analyses in subgroups formed by dividing groups according to culture. This was done since culture emerged as a significant moderator even when publication bias was modelled. PET-PEESE indicated significant publication bias in the North-American dataset (see Table 3). This is in line with a visual inspection of the funnel plot (see Figure 2). All PET-PEESE adjusted effect sizes were substantially smaller than the unadjusted effect size and no longer significantly different from zero. The 3-PSM did not indicate significant fit improvement after publication bias was modelled but

also led to a smaller corrected effect size that was no longer significantly different from zero in the European dataset (see Table 3). Given that there was only one study on a South-East Asian sample, neither PET-PEESE nor the 3-PSM could be implemented in that dataset.

Exploratory analyses

Several reasons might explain the diverging results from PET-PEESE and 3-PSM (see Discussion), but one of them was further explored empirically in exploratory follow-up analyses. These were not preplanned and have to be interpreted as such. We investigated how the selection threshold affects the results of the 3-PSM. Specifically, we picked other values than the two-tailed $p = .05$ (one-tailed $p = .025$ in the expected direction) as the threshold below which studies might have been more likely to be selected for publication than those above it. Higher thresholds might be plausible given the use of one-tailed tests or the regard of values above but close to $p = .05$ as ("marginally") significant (for further clarification see Discussion). The results indicated that the overall effect size was reduced to approximately $d = 0.05$ or lower for one-tailed selection thresholds between $p = .05$ and $p = .15$, corresponding to two-tailed p- values of .10 and .30, respectively (see Figure 3). The corresponding likelihood ratio tests appear to indicate that this is accompanied by beyond chance fit-improvement, with p-values below .001 in the corresponding interval. Regarding the potential type-1-error accumulation inherent in this analysis, it should be noted that the lowest p-value remains significant after Bonferroni correction for the (arbitrary, $n = 981$) number of tests displayed, $p\_adjusted < .001$. Nevertheless, this has to be interpreted with caution given the strictly exploratory nature of these analyses.

Discussion

Overall, both PET-PEESE and 3-PSM led to reductions in the average moral licensing effect size. Specifically, PET-PEESE reduced the effect size to $d = -0.05$ and the 3-PSM reduced it to $d = 0.18$. These estimates are substantially smaller than previous meta-analytic estimates of $d = 0.31$ and $d = 0.32$ (Blanken et al., 2015; Simbrunner & Schlegelmilch, 2017, respectively). Furthermore, PET-PEESE indicated significant publication bias, as supported by an asymmetric funnel plot. The 3-PSM, however, did not indicate significant fit improvement after publication bias was modelled.

In subsamples built by culture, PET-PEESE reduced the effect sizes from North-American and European samples to $d = -0.13$ and $d = 0.10$, respectively. While the 3-PSM somewhat converged in the European sample, $d = 0.11$, the effect size in the North American sample was reduced to a far lesser degree, $d = 0.31$. Overall, the findings from PET-PEESE and 3-PSM do point in a similar direction, i.e., that the moral licensing effect size has been overestimated due to publication bias. Nevertheless, they show substantial differences that require further consideration.

Comparison PET-PEESE and 3-PSM

Statistical power and uncertain parameter estimates. To some degree, the diverging results from PET-PEESE and the 3-PSM might be attributable to chance. For instance, it could be argued that the statistical power to find evidence for publication bias was not high enough, leading to the nonsignificant result of the 3-PSM even in the presence of selection for significance. Regarding the corrected effect size estimates, the results are very similar in the European sample with $d = 0.10$ and $d = 0.11$ using PET-PEESE and the 3-PSM, respectively.

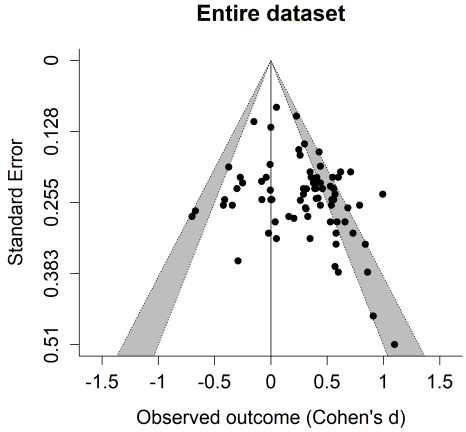


Figure 1. Contour-enhanced funnel plot of the entire dataset with a reference line at $d = 0$. Effect sizes within the white triangle are not significantly different from zero at the .05 level. The grey area begins at $p = .05$ and ends at $p = .01$. Therefore, effect sizes that surpass the second grey border are significant at the .01 level.

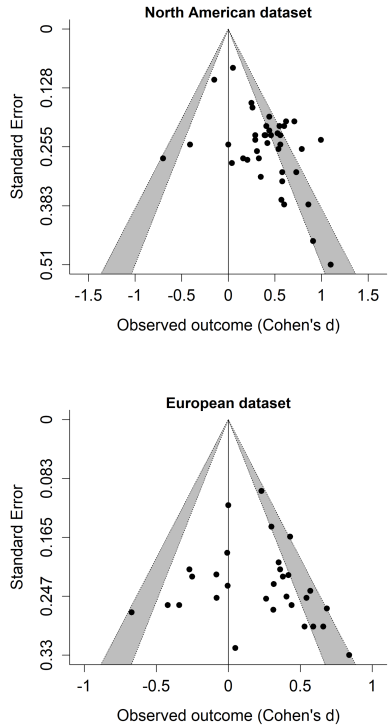
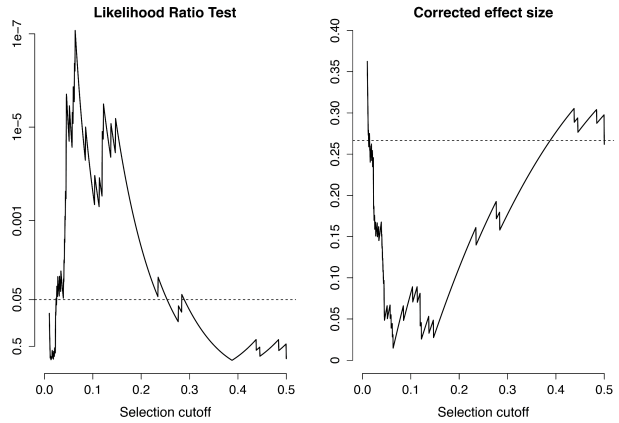


Figure 2. Contour-enhanced funnel plots of the North American and the European datasets with a reference line at $d = 0$. Effect sizes within the white triangle are not significantly different from zero at the .05 level. The grey area begins at $p = .05$ and ends at $p = .01$. Therefore, effect sizes that surpass the second grey border are significant at the .01 level.

In the North American sample however, the results clearly diverge, with d = -0.13 and d = 0.31 from PET-PEESE and the 3-PSM, respectively. This is unlikely to be attributable to random variation given that the 95%-confidence intervals around the corrected effect sizes do not overlap. Therefore, these differences require further explanation. In the overall sample, the effect sizes are also dissimilar, with $d = -0.05$ and $d = 0.18$, respectively. However, the 95%-confidence intervals show some overlap. Specifically, effect sizes between $d = 0.06$ and $d = 0.16$ are part of both confidence intervals and would hence be consistent with the results from both methods.

Limitations of PET-PEESE. A potential criticism of PET-PEESE is that relationships between effect size and standard error can stem from benign causes such as power calculations. If researchers correctly anticipated their effect sizes and adjust their sample size to it, PET-PEESE would have corrected the effect size downwards in the absence of publication bias. However, moderators of the moral licensing effect are largely unknown (Blanken et al., 2015). Furthermore, the significant relationship between the effect sizes and their standard errors persisted in a model that included culture and type of comparison - two moderators suggested by Simbrunner and Schlegelmilch (2017). These findings speak against the proposition that researchers accurately anticipated the effect sizes of their studies (compared to other moral licensing studies) and used these estimates in power-analyses. In line with this, several large studies represent exact or slightly modified replications of initial smaller studies, leading to substantially reduced effect sizes in the replications (Blanken et al., 2014; Ebersole et al., 2016; Urban et al., 2017). Therefore, the significant slope of PET-PEESE can be most parsimoniously explained as resulting from publication bias.

Figure 3. Plotted is the fit improvement (p-value of likelihood ratio test) and the corrected effect size for 3-PSMs



using different selection cutoffs. The selection cutoffs plotted on the x-axis represent one-tailed p-values. The dashed lines indicate $p = .05$ (left plot) and the uncorrected meta-analytic effect size (right plot), respectively.

Another limitation of PET-PEESE that might explain some of the inconsistencies with the 3-PSM is its suboptimal performance under certain conditions as indicated by simulation studies (Carter et al., 2018; Stanley, 2017). For instance, the simulations from Carter et al. (2018) indicated that PET-PEESE can lead to overestimations of an effect size of zero in the presence of large heterogeneity and publication bias. Importantly however, it can also lead to underestimations of an existing effect, especially in the presence of questionable research practices – although this property is shared to some degree by

the 3-PSM. Therefore, it cannot be excluded that the correction of the overall effect size to d = -0.05 represents an over-correction while the estimates from the 3-PSM could be biased slightly upwards due to reasons discussed below.

Limitations of the 3-PSM. One issue of the 3-PSM is the common phenomenon that studies often report data on multiple outcomes per study (see Blanken et al., 2015, Table 1). For instance, in a study by Mazar and Zhong (2010, study 3) participants were given the opportunity to both lie about their task performance to gain additional money and to steal money in a self-gratification paradigm. However, selection for significance might only occur or predominantly occur for one focal effect size that cannot be determined with certainty by meta-analysts. In these cases, a common practice that was also taken here is to average over multiple effect sizes. However, selection for significance never occurred based on this average in practice. We simulated a scenario in which averages over two dependent variables – only one of which was used for selection – were used in a 3-PSM. As expected, our simulations showed that that the 3-PSM did not always recover the true effect size in these cases. Specifically, publication bias led to overestimations of the true effect size that was no longer corrected downward to a sufficient degree by the 3-PSM (see Appendix). Given that we also had to average over multiple effect sizes in the present dataset, this problem might have occurred in our study. This might have led to an overestimation of the true effect size by the 3-PSM.

A further issue with the 3-PSM is the choice of a selection threshold. While a two-tailed $p < .05$ might be the most frequently employed criterion to judge statistical significance, studies could also be selected for publication as long as their p-values are sufficiently close to this value. Specifically, p-values between .05 and .10 are often considered "marginally significant" and are often interpreted as constituting at least some support for the hypothesis. This also occurred in the moral licensing literature (e.g., Effron et al. 2012; Jordan et al. 2011; Clot, Grolleau, & Ibanez, 2014). Furthermore, some authors might use one-tailed hypothesis tests if they made a prediction about the direction of the effect. Again, this also occurred in the moral licensing literature (e.g., Conway & Peetz, 2012; Susewind, & Hoelzl, 2014). In these two cases, the selection threshold below which studies are more likely to be published appears to be above a one-tailed p = .025 and perhaps closer to a one-tailed p = .05 or even higher. In order to explore the impact of the choice of selection threshold on the results of the 3-PSM, we repeated the analyses with several different thresholds. Our results indicated that the 3-PSM reduced the overall effect size to almost zero for selection thresholds between one-tailed p = .05 and p = .15, which would have been accompanied by fit improvement after selection was modelled. While this finding does not appear to be attributable to chance, it has to be interpreted with caution given that the analysis was strictly exploratory and only carried out to explore the diverging results of PET-PEESE and the 3-PSM. Nevertheless, this hints at the possibility that common practices of considering findings with one-tailed p-values above .025 as hypothesis-confirming has led the 3-PSM to overestimate the true effect size when using this threshold.

Interim Conclusion. Overall, PET-PEESE and the 3-PSM have limitations that can lead to them both under- and overestimating the true effect size. Our findings provide an example of the difficulties associated with the adjustment for publication bias in practice. In the present study, the 3-PSM estimates appeared to be substantially larger than those from PET-PEESE. This might to some degree be attributable to the fact that PET-PEESE does in some cases underestimate a true effect size (Carter et al., 2018). However, there are also two plausible reasons suggesting that the 3-PSM might have overestimated the true effect size: (1) Its impaired performance when focal and non-focal effect sizes are averaged for effect size calculation and (2) the issue that the "true" selection threshold might occasionally be above one-tailed p = .025 in practice. In sum, it cannot be stated with certainty that the true effect size of moral licensing is essentially zero – even the d = -0.05 estimate from PET-PEESE has an upper 95% CI boundary of d = 0.16. Nevertheless, there is good reason to conclude that the average moral licensing effect is very small and that even the d = 0.18 estimate from the 3-PSM might still be an overestimation.

Moderator effects

Regarding the role of culture, we could only partially confirm the findings from Simbrunner and Schlegelmilch (2017). Specifically, our results only

indicated a significant difference between North American and South-East Asian samples. The difference between European and South-East Asian samples was no longer significant. The findings diverge from those reported by Simbrunner and Schlegelmilch (2017) because they included multiple outcomes from the same study as separate effect sizes while we decided to average over outcomes to satisfy the independence assumption of classical meta-analysis. In general, the facts that there is only one study from South-East Asia and that publication bias probably exaggerated the moral licensing effect size in other cultures (see above) render the evidence for an attenuated or reversed effect in South-East Asia weak. However, the difference between North American and European studies was also significant, even after publication bias was modelled. This suggests that the moral licensing effect might be of larger magnitude in North America, which deserves further investigation (see Recommendations for further research).

For the comparison between moral licensing effects with neutral and immoral control conditions, our findings also indicated a significant difference, although Simbrunner and Schlegelmilch (2017) reported a much larger effect. This could largely be attributed to the accidental inclusion of two effect sizes as $d > 3$ in the immoral condition. Furthermore, the moderator was no longer significant after publication bias was modelled. Therefore, the evidence for this moderator is at present not very robust. It should be noted that an immoral control condition also conflates moral licensing with moral cleansing and might lead to significant group differences in the absence of licensing effects. Nevertheless, this moderator also deserves further investigation (see Recommendations for further research).

Recommendations for further research

If researchers seek to firmly establish the existence of the classical moral licensing effect, more large studies such as the Ebersole et al. (2016) multi-lab project should be pursued. In order to prevent QRPs and publication bias, they should be pre-registered. Using the more optimistic effect size estimate from the 3-PSM of $d = 0.18$, a minimum of $n = 766$ participants would be required for 80% power in a one-tailed test. In order to ensure that null results are not attributable to failed activations of the moral self-concept, subsequent studies should use previously validated manipulations to ensure a valid test of moral licensing.

Future meta-analytic work on the moral licensing effect would benefit from the use of three-level meta-analysis (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). Three-level meta-analysis allows the inclusion of multiple dependent effect sizes from the same sample as is common in the moral licensing literature (e.g., due to multiple outcomes or multiple experimental conditions compared to the same control condition). Using this three-level framework, it is feasible to perform moderator-analyses even when the moderator in question varies within studies (López-López, Van den Noortgate, Tanner-Smith, Wilson, & Lipsey, 2017). It could therefore be valuable for the identification of moderators of the moral licensing effect from existing literature.

Regarding future empirical work aiming to identify potential moderators of moral licensing, research should continue to examine the role of culture. This moderator might be of particular interest given that the contrast between North American and European samples remained significant even after publication bias was modelled. However, up to now there is only one other sample from different parts of the world (i.e., South-East Asia). Further research on moral licensing should therefore cover a larger variety of different cultural backgrounds in order to investigate the moderating role of culture in more detail. To this end, instead of coding broad world regions, a more differentiated way of coding cultural categories in both existing and new studies could be applied. For instance, participants' cultural background could be coded according to the individualism-collectivism dimension (e.g., by using the respective country average values provided by Hofstede, 2001). The individualism-collectivism dimension captures cross-cultural differences in interdependent versus independent self-construal (Triandis, 2018). Given the importance of the moral self-concept for moral licensing (Sachdeva et al., 2009) and that differences in moral values are related to the individualism-collectivism dimension (Triandis, 2001), this dimension could be a promising moderator. However, it should also be noted that cultural differences might not have a moderating impact on moral licensing in general, but only on the conditions under which the effect might occur. In general,

future empirical but also theoretical work is required to understand the potential role of culture in moral licensing.

Finally, differential moral licensing effects depending on the applied comparison type (i.e., neutral versus immoral behavior) deserve further investigation. We replicated the finding of immoral control conditions leading to larger moral licensing effect sizes than neutral control conditions. Nevertheless, this effect disappeared after publication bias was modelled and was substantially less robust than reported by Simbrunner and Schlegelmilch (2017). Comparing a moral licensing condition with both a neutral and an immoral control condition might be of particular theoretical interest given that the moral self-regulation framework (Sachdeva et al., 2009) introduces moral cleansing (i.e., immoral behavior leading to subsequent moral behavior) as a counterpart of moral licensing. However, it cannot be recommended that future studies seeking to demonstrate the moral licensing effect use only an immoral control condition. In these cases, potential licensing effects might instead be attributable to cleansing effects. Future meta-analytical work should simultaneously assess the evidence for moral licensing and for moral cleansing.

Conclusion

Overall, we found some evidence for publication bias in the moral licensing literature. The mean effect sizes were very small to small after the correction for publication bias and might in part still be exaggerated. Large samples would be required to allow for a meaningful study of these effects - sample sizes that were not achieved in the original studies on moral licensing. We found some evidence that culture moderated the moral licensing effect, which should be further explored in subsequent studies. Most importantly, future research should aim for high power and employ pre-registration.

Open Science Practices

This article earned the Open Data and the Open Materials badge for making the data and materials available: https://osf.io/reky2. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement. More information about the badges: https://osf.io/tvyxz/wiki/home/

References

References marked with an asterisk are exclusively cited in Table 1 or Table 2.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource?. *Journal of personality and social psychology*, 74, 1252-1265.

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111-125). Chichester, England: John Wiley & Sons.

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, 6, 1-62.

*Blanken, I., Van de Ven, N., & Zeelenberg, M. (2012). [A variety of studies on the self-licensing effect]. Unpublished raw data.

Blanken, I., van de Ven, N., Zeelenberg, M., & Meijers, M. H. (2014). Three attempts to replicate the moral licensing effect. *Social Psychology*, 45, 232-238.

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41, 540-558.

*Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., & McKenna, C. (2010). Moral credentialing by association: The importance of choice and relationship closeness. *Personality and Social Psychology Bulletin*, 36, 1564-1575.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated?. *Frontiers in psychology*, 5, 823.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2018, February 1). Correcting for bias in psychology: A comparison of meta-analytic methods. Retrieved from: http://doi.org/10.17605/OSF.IO/9H3NU

Clot, S., Grolleau, G., & Ibanez, L. (2014). Smug alert! Exploring self-licensing behavior in a cheating game. *Economics Letters*, 123, 191-194.

Coburn, K. M., & Vevea, J. L. (2017). weightr: Estimating weight-function models for publication bias in R. R package version 1.1.2.

Conway, P., & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. *Personality and Social Psychology Bulletin*, 38, 907-919.

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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82.

*Effron, D. A. (2014). Making mountains of morality from molehills of virtue: Threat causes people to overestimate their moral credentials. *Personality and Social Psychology Bulletin*, 40, 972-985.

*Effron, D. A., Monin, B., & Miller, D. T. (2012). Inventing racist roads not taken: The licensing effect of immoral counterfactual behaviors. *Journal of Personality and Social Psychology*, 103, 916-932.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315, 629-634.

Etherton, J. L., Osborne, R., Stephenson, K., Grace, M., Jones, C., & De Nadai, A. (2018). Bayesian analysis of multimethod ego-depletion studies favours the null hypothesis. *British Journal of Social Psychology*, 57, 367-385.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7, 555-561.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, 19, 975-991.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345, 1502-1505.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological bulletin*, 82, 1-20.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... & Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546-573.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, 136, 495-525.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, 9, 61-85.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, 21, 299-332.

Hofstede, G. J. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed). Thousand Oaks, CA: Sage publications.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18, 235-241.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109-117.

*Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, 37, 701-713.

*Kouchaki, M. (2011). Vicarious moral licensing: The influence of others' past moral actions on moral behavior. *Journal of personality and social psychology*, 101, 702-715.

López-López, J. A., Van den Noortgate, W., Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2017). Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation. *Research synthesis methods*, 8, 435-450.

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people?. *Psychological science*, 21, 494-498.

*Meijers, M. H. C., Noordewier, M. K., Verlegh, P. W. J., & Smit, E. G. (2014). *Identity relevance moderates the licensing effect* (Chapter from doctoral dissertation, Amsterdam School of Communication Research). Retrieved from http://dare.uva.nl/record/1/432499

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and personality psychology compass*, 4, 344-357.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of personality and social psychology*, 81, 33-43.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC medical research methodology*, 9, 2.

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

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

R Core Team (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: John Wiley & Sons.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological science*, 20, 523-528.

Scargle, J. D. (2000). Publication Bias: The "File-Drawer" Problem in Scientific Inference. *Journal of Scientific Exploration*, 14, 91–106.

Schonemann, P. H., & Scargle, J. D. (2008). A Generalized Publication Bias Model. *Chinese Journal of Psychology*, 50, 21–29.

*Simbrunner, P., & Schlegelmilch, B. B. (2016). [A variety of studies on the moral licensing effect]. Unpublished raw data.

Simbrunner, P., & Schlegelmilch, B. B. (2017). Moral licensing: a culture-moderated meta-analysis. *Management Review Quarterly*, 67, 201-225.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22, 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534-547.

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8, 581-591.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5, 60-78

Susewind, M., & Hoelzl, E. (2014). A matter of perspective: Why past moral behavior can sometimes encourage and other times discourage future moral striving. *Journal of Applied Social Psychology*, 44, 201-209.

Triandis, H. C. (2018). Introduction: Two constructs. In R. Nisbett (Ed.), *Individualism and collectivism* (pp.1-18). New York, NY: Routledge.

Triandis, H. C. (2001). Individualism-collectivism and personality. *Journal of personality*, 69, 907-924.

Urban, J., Bahník, Š., & Kohlová, M. B. (2017, December 19). Green consumption does not make people cheat: Three replications of a moral licensing experiment. http://doi.org/10.17605/OSF.IO/WYNJB

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior research methods*, 45, 576-594.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60, 419-435.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1-48.

*Young, Y., Chakroff, A., & Tom, J. (2012). Doing good leads to more good: The reinforcing power of a moral self-concept. *Review of Philosophy and Psychology*, 3, 325-334.

Zhong, C., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, 313, 1451–1452.

Appendix

The influence of averaging over outcomes on the performance of 3-PSM

In order to investigate how averaging over different outcomes per study affects the performance of the 3-PSM, a small simulation study was performed. Specifically, we simulated three meta-analytic scenarios: (1) The ideal case: Only the dependent variable based on which selection for significance (publication bias) occurred was used for effect size calculation. (2) In all cases, selection occurred based on one dependent variable, but the effect size was always calculated by averaging over this dependent variable and a different outcome that correlates r = .5 with the former. (3) 50% of the effect sizes corresponded to the ideal case, 50% to the second scenario. These scenarios were simulated for combinations of: (1) different heterogeneities: I^2 = 0, 0.1, 0.2, and 0.3, (2) different selection probabilities given that a study is nonsignificant: 1, 0.5, 0.25, and 0.1 as well as (3) different true effect sizes: 0, 0.1, 0.2, 0.3, 0.4, and 0.5 (standardized mean difference). For each scenario, a "naïve" random effects meta-analysis not correcting for publication bias and a 3-PSM was calculated. Each scenario was simulated 1000 times. For certain effect size × heterogeneity × publication bias combinations, no results are presented since too few studies were significant to compute a 3-PSM in several of the iterations.

The plots below indicate the degree of bias for the different conditions. As can be seen, increasing degrees of publication bias, especially when combined with increasing heterogeneity, lead to an overestimation of the true effect size. This bias is strongest for small effect sizes. Furthermore, the 3-PSM tends to recover the true effect size on average in the ideal case where only the dependent variable that selection was based on was used for effect size calculation. By contrast, the performance of the 3-PSM was not optimal when 50% and even more so when 100% of the effect sizes were based on averages over multiple dependent variables (only one of which was used for selection). In these cases, the 3-PSM overestimated the true effect size. Again, this overestimation was strongest for small true effect sizes. It was increased by the degree of publication bias and heterogeneity. In the most extreme case, the bias of the 3-PSM-adjusted estimate exceeded d = .10.

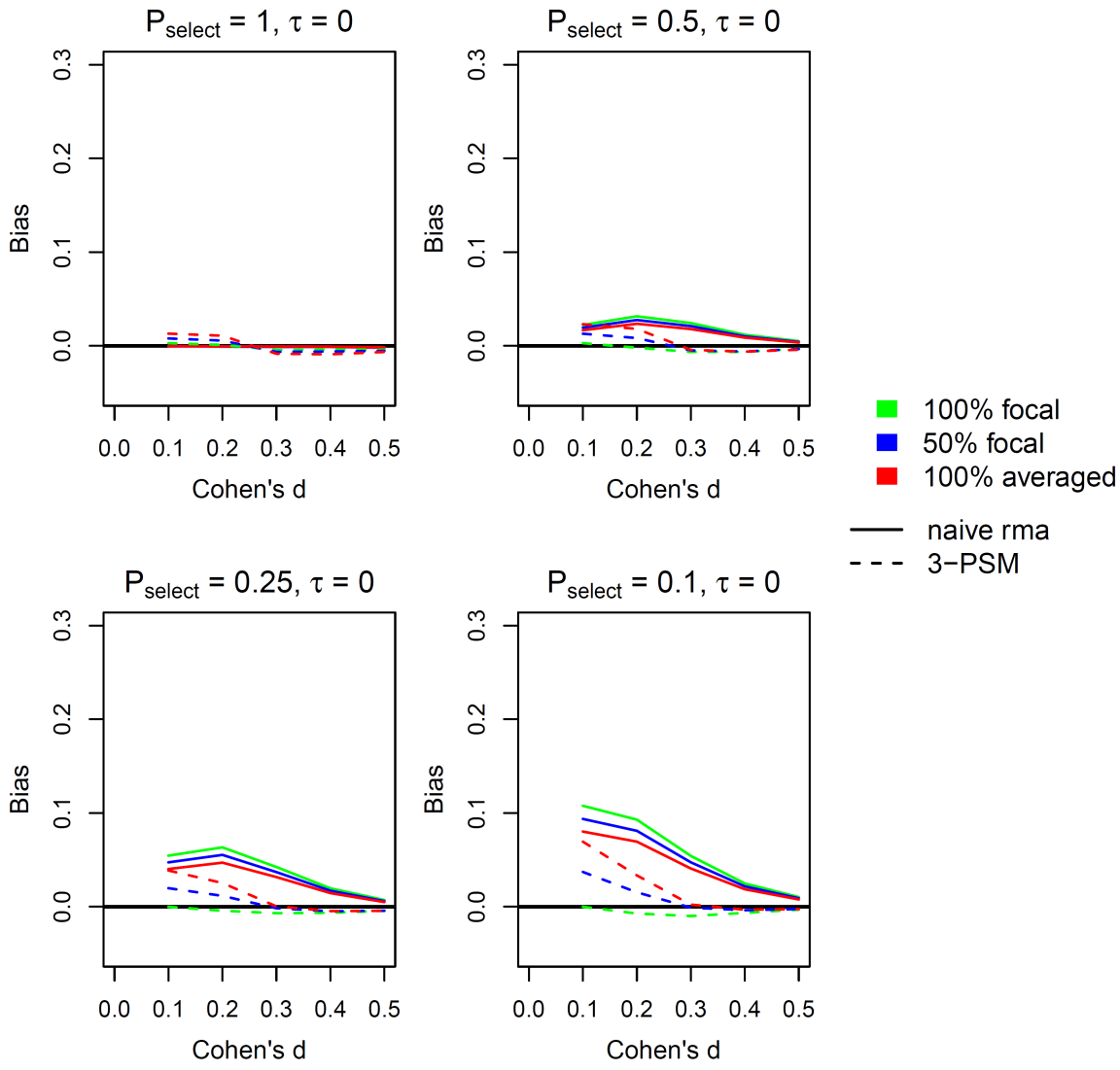


Figure A1. Displayed is the average bias of the effect size estimates for different scenarios. The true effect size is displayed on the x-axis. P_{select} = Probability that a nonsignificant study is selected for publication, indicating the degree of publication bias. τ = Between-study heterogeneity. Continuous lines indicate uncorrected meta-analytic estimates, dashed lines 3-PSM adjusted estimates. 100% focal = effect sizes are based entirely on the dependent variable used for selection. 100% averaged = effect sizes are based on averages of the dependent variable used for selection and another correlated dependent variable. 50% focal = half of the effect sizes stem from each of the prior scenarios.

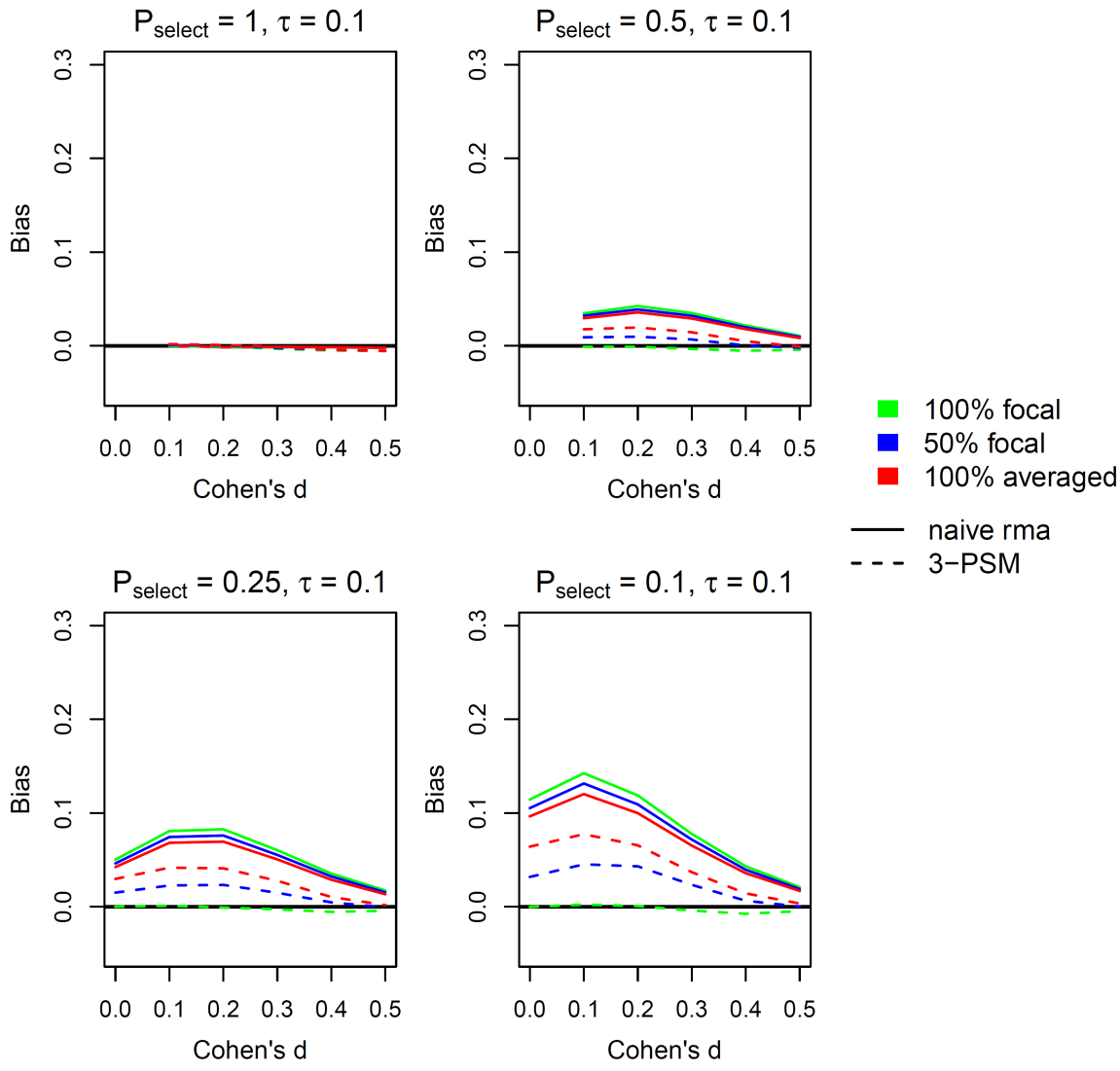


Figure A1. (continued)

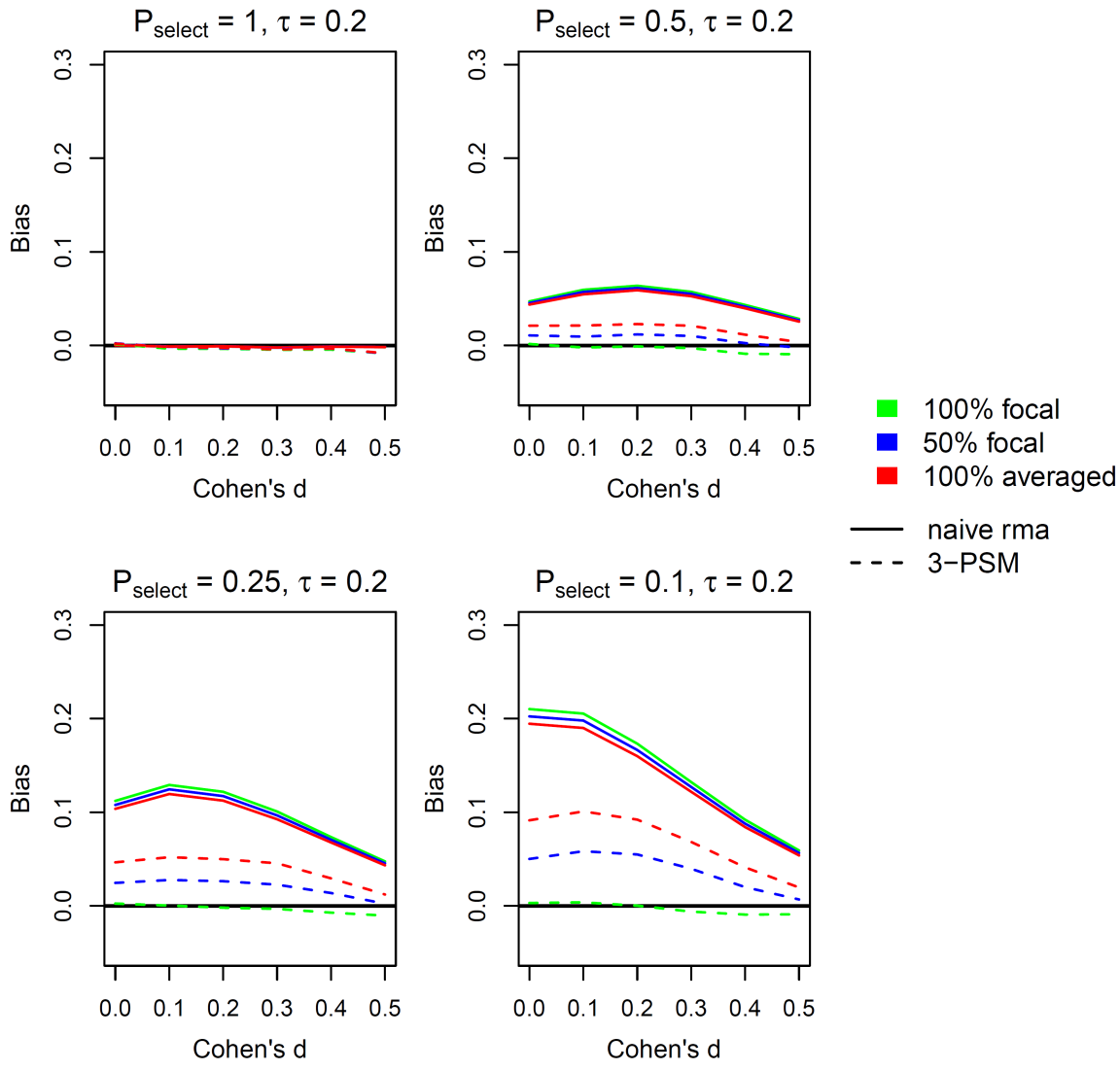


Figure A1. (continued)

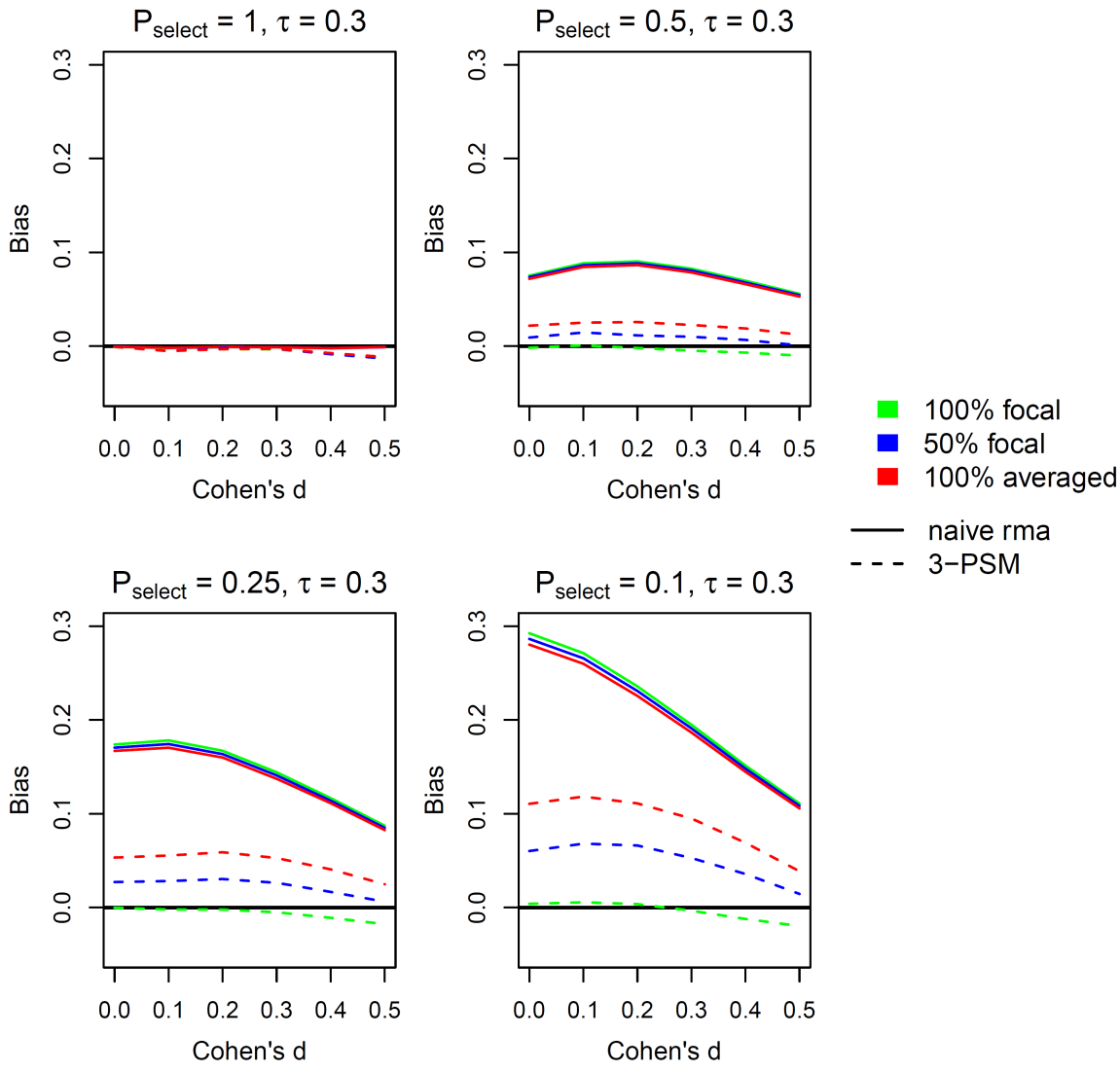


Figure A1. (continued)