

Influences of Past Moral Behavior on Future Behavior: A Review of Sequential Moral Behavior Studies Using Meta-Analytic Techniques

Participant-visible method context. Results, discussion, captions, and the target figure are excluded.

Audience for Manipulation and Target Behavior In some SMB studies, efforts are made to make moral manipulation and target behavior appear unrelated. This may be by ostensibly presenting them as part of two different studies being conducted for separate research teams, but with a shared data collection session (e.g., Jordan et al., 2011), or by having each task presented by two seemingly unrelated people (e.g., Dolinski, 2012). In essence, this represents an attempt to make the tasks appear as though they are being observed by two distinct audiences. In most studies, the relationship between the manipulation and target behavior is not disclosed, but no effort is made to make them appear as though they are being observed by different audiences. Existing theories do not make explicit predictions about whether having the same or different audiences for the manipulation and target behavior should impact SMB effects. However, having the same audience may introduce a source of hypocrisy concerns, which the moral credits model suggests may suppress licensing (Effron & Monin, 2010). Evidence of the effect of audience is mixed. On the one hand, Monin and Miller (2001) found that moral licensing occurred both in conditions in which the moral manipulation and target behavior were observed by the same audience and conditions in which they were observed by ostensibly unrelated audiences. On the other hand, Kristofferson et al. (2014) found moral licensing only occurred if both the manipulation and target behavior were performed publicly (i.e., observed by the same audience) but found positive consistency occurred when the manipulation involved behaviors that were performed privately (i.e., not observed by the same audience as the target behavior). Greene and Low (2014) found that moral licensing only occurred when the target behavior (engagement in hypothetical moral transgressions) was described to participants as being private (i.e., not observed by others). We examined whether efforts to make participants believe the moral manipulation and target behavior were observed by unrelated audiences would moderate effect sizes. We did not make specific predictions about how the audience would moderate effect sizes due to both the lack of theoretical predictions regarding this and the inconsistency in the results of previous literature.

Year of Publication Prosocial-FITD research dates to the 1960s (e.g., Freedman & Fraser, 1966) and typically reports evidence of positive consistency. The concept of moral balancing was not formally discussed until much later (Nisan, 1990) and did not come to prominence until the early 2000s with the publication of Monin and Miller's (2001) research. Due to the disparity in when these effects emerged, we chose to examine whether effect sizes were moderated by publication year. Changes in results over time could reflect changes in social norms relating to moral behavior, changes in research practices (e.g., greater emphasis on larger sample sizes and statistical power), or merely a change in research interest from using the specific variation of SMB design seen in prosocial-FITD literature to more novel variations seen in moral licensing literature. We also examined whether there were

Transparency and Openness We did not preregister a protocol for these meta-analyses. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses reporting guidelines for the final report. Additional online materials can be found at <https://osf.io/upby2/>. These additional online material include a full list of search terms with comparisons to previous meta-analyses (Table S1: <https://osf.io/yb589/>), a list of studies excluded during full-text screening with exclusion reason (Table S2: <https://osf.io/yb589/>), and effect size information (study characteristics, effect size data, moderator coding) for both Meta-Analysis 1 (Table S3: <https://osf.io/yb589/>) and Meta-Analysis 2 (Table S5: <https://osf.io/yb589/>). Data files and the analysis script can also be found via this link (<https://osf.io/upby2/>).

Literature Search We conducted a systematic search of the following databases: APA PsycInfo, Proquest, Medline, and Web of Science Core Collection.² The search was conducted to identify studies for both Meta-Analysis 1 (positive manipulations) and Meta-Analysis 2 (negative manipulations) concurrently and so it included terms relevant to both. The search was originally conducted in April 2015 and updated in January 2019, November 2020, and July 2022. SMB studies come from a range of literature that use different terms to describe research designs and effects, as the term SMB paradigm was not coined until relatively recently (Mullen & Monin, 2016). As such, we used multiple search strings to target different areas of literature known to publish SMB effects. The first string targeted moral balancing literature and combined terms for three concepts: (a) morality (e.g.,

moral*), (b) observed effects (e.g., licens*), and (c) outcome measures of moral and immoral behavior (e.g., help*). The second search string targeted the sequential request and guilt and compliance literature and included terms relating to two concepts: (a) effect terms relating to FITD, guilt and compliance, and door-in-the-face (DITF) literature (e.g., FITD) and (b) moral behavior (e.g., prosocial). The third search targeted the environmental spillover literature and included terms relating to two concepts (a) spillover (e.g., spillover OR spill-over) and (b) environmental behavior (e.g., pro-environment*). All search terms can be found in the additional online material (Table S1: <https://osf.io/yb589>). Although these terms are not exhaustive, they were adequate to capture a broad representation of SMB studies. Additional online material (Table S1: <https://osf.io/yb589>) also provides a comparison of search terms used in previously published

At the time of the searches, the institutional subscription to Web of Science covered Science Citation Index Expanded, Social Sciences Citation Index, Arts and Humanities Citation Index, Conference Proceedings Citation Index—Science, Conference Proceedings Citation Index—Social Science and Humanities, Book Citation Index—Science, Book Citation Index— Social Sciences and Humanities, Emerging Sources Citation Index, Current Chemical Reactions, and Index Chemicus.

FERGUSON, KAUFMANN, BROWN, AND DE LA PIEDAD GARCIA

behavior over time were excluded to avoid the possibility that intervening events may confound the results. This reasoning is grounded in assumptions of the moral credits model, which proposes each instance of engaging in a morally salient behavior has the potential to change moral self-image and thus future moral behavior. If there is a substantial temporal gap between the manipulation and target behavior, there is an opportunity for participants to engage in additional morally salient actions that could alter their moral self-image and contaminate the results. This means it would not be possible to determine if the participant's behavior is due to the moral manipulation or due to other actions that participants have engaged in during the intervening time.

meta-analyses on related topics (e.g., moral licensing, FITD). We also searched reference lists of included studies and review articles and meta-analysis on related topics (e.g., Blanken et al., 2014; Geiger et al., 2021). In addition to the database search, calls for unpublished data were distributed through the list serves and online forums of the Australasian Society for Social Psychology, the Society for Personality and Social Psychology, and the European Association for Social Psychology. We also contacted prominent authors in the field directly to enquire about unpublished data.

Study Selection Screening for both Meta-Analyses 1 and 2 was conducted concurrently. First, the titles and abstracts of all records located through the search strategy were screened for relevance. Those selected for full-text screening were screened by the first author, and a subset of 110 records were double-screened by other members of the research team. Any disagreements were resolved through discussion with a third screener. The inclusion criteria for Meta-Analysis 1 were as follows: 1. Participants had to be individuals. Studies measuring household behavior (e.g., household resource use) or decisions by organizations were excluded. 2. Studies had to use a between-groups, experimental design using the SMB paradigm, with random allocation to experimental conditions (correlational studies or studies in which participants were grouped based on initial behavior, rather than it being manipulated, were excluded). 3. Studies had to include a positive moral manipulation based on actual, recalled, or imagined behavior. We also included SMB studies using trait-based manipulations (e.g., writing about having positive moral traits; Sachdeva et al., 2009). These manipulations are argued to directly impact participants' moral self-image and have been used in studies that have found both moral licensing (Sachdeva et al., 2009) and positive consistency (Aquino et al., 2009). Studies that used manipulations based on nonmoral behavior or other people's behavior were excluded. 4. Studies had to include a control condition in which participants did not perform, imagine, or recall their own moral behavior. We included studies in which participants in the control condition either (a) performed, recalled, or imagined a morally neutral behavior, (b) observed, recalled, or imagined someone else's behavior, or (c) were passive controls (performed no manipulation tasks). We excluded studies that compared positive manipulation conditions to negative manipulation conditions (i.e., no control group) to avoid ambiguity in interpreting results (for further discussion on the need for neutral control conditions in SMB research, see Mullen & Monin, 2016). 5. Studies had to include a morally positive or negative target behavior that was measured in the same experimental session as the moral manipulation. Studies in which the moral manipulation and target behavior were measured in separate sessions, or longitudinal studies tracking

Reports had to be written in English or have verified English translations available. This was to avoid any errors resulting from incorrect translation by the research team. Only two reports were excluded for this reason.

Data Extraction The first author completed data extraction for all included studies. Approximately 41% of the included studies ($k = 91$) were doubleextracted by the third or fourth author. Discrepancies in data extraction were minimal (two effect sizes differed between coders due to an error in extracting cell sizes; these discrepancies were resolved by referring to the study reports). Interrater agreement was high for coding of moderators (see Table 2). Due to the high level of agreement between extractors and due to practical constraints, it was decided that the remaining effect sizes would be extracted by the first author only.

Study Characteristics. For each study, we recorded a description of the moral manipulation, control condition, and target behavior (see additional online material, Table S2: <https://osf.io/yb589>). If multiple target behaviors were included in a study, we extracted data from the first measure presented to participants when the presentation occurred in a fixed order. If the order of presentation was randomized ($k = 4$ effect sizes), we averaged these together into a single effect size.

Moderator Coding. Each moderator category was coded as described below.

Target Behavior Valence. Studies were coded to identify whether they included a positive (e.g., charitable donations) or negative target behavior (e.g., cheating).

Domain-Consistency. Studies were coded as domain-consistent if the manipulation and target fit into the same broad conceptual category (e.g., pro-environmental behavior; racial prejudice; sharing). They were coded as domain-inconsistent if the manipulation and target did not fit the same broad category. Where the domain of the manipulation was not possible to identify (e.g., recall manipulations that just ask participants to recall acting morally; Stellar & Willer, 2014), the studies were coded as uncategorized and omitted from the analysis of this moderator.

Manipulation Behavior Type. Manipulations were classified into one of three categories: (a) performed behavior (e.g., signing a petition; Kristofferson et al., 2014), (b) recalled behavior, or (c) hypothetical behavior, such as responding to hypothetical choices (e.g., Berger et al., 2020), agreeing to perform an action in the future (Cascio & Plant, 2015), or writing about themselves using moral

REVIEW OF SEQUENTIAL MORAL BEHAVIOR

Table 2 Descriptive Statistics for Positive Sequential Moral Behavior Studies, Meta-Analysis 1 Sample size

Publication year

Moderator

k

M

SD

Mdn

Range

M

SD

Mdn

Range

All studies Target behavior valence ($\kappa = .98$) Negative Positive Domain-consistency ($\kappa = .82$) Inconsistent Consistent Manipulation behavior type ($\kappa = .79$) Hypothetical Recall Behavioral Manipulation focus ($\kappa = .83$) Identity-focused Behavior-focused Target behavior type ($\kappa = .80$) Hypothetical Behavior Area of literature ($\kappa = .92$) Non-FITD Prosocial-FITD Audience ($\kappa = .75$) One audience Two audiences Publication status ($\kappa = 1.00$) Unpublished Published

143.97

248.02

84.00

16–3,286
2011.63
8.27
1972–2022
159.97 136.83
402.99 130.65
80.00 84.50
31–3,286 16–708
2013.54 2011.23
4.61 9.45
2001–2021 1972–2022
158.66 156.30
154.38 326.96
86.00 86.00
29–687 31–3,286
2012.40 2010.05
7.79 10.20
1981–2021 1972–2022
159.11 139.67 128.20
351.43 148.05 110.40
84.00 76.00 84.00
16–3,286 29–708 29–686
2013.00 2013.15 2008.12
4.44 11.79 12.13
2001–2020 2008–2022 1972–2021
110.16 155.25
119.11 281.72
67.00 86.00
16–567 29–3,286
2012.2 2011.30
3.75 9.49
2001–2020 1972–2022
163.30 125.86
337.39 111.17

82.00 88.00
 29–3,286 16–686
 2011.50 2011.76
 8.09 8.47
 1975–2022 1972–2021
 157.42 89.78
 272.80 49.90
 85.00 74.00
 16–3,286 29–229
 2013.77 2001.29
 4.21 13.85
 2001–2022 1972–2018
 170.61 74.39
 301.72 39.72
 93.00 64.00
 18–3,286 16–157
 2011.87 2008.07
 8.71 11.67
 1972–2022 1977–2021
 116.93 153.84
 100.45 282.99
 80.50 84.00
 29–447 16–3,286
 2013.25 2010.80
 2.54 9.42
 2010–2020 1972–2022

Note. Outliers are omitted from Meta-Analysis 1. FITD = foot-in-the-door.

trait words (Sachdeva et al., 2009). Manipulations that included elements of multiple categories (e.g., recalling previous behavior and writing about having moral traits; Blanken et al., 2012) were omitted from moderation analyses. Manipulation Focus. Manipulations were coded dichotomously as either behavior-focused or identity-focused according to criteria identified by Effron and Conway (2015). Specifically, behavior-focused manipulations were those that (a) were accompanied by instructions or primes to think about the behavior concretely (e.g., to focus on immediate consequences; Cornelissen et al., 2013), (b) were accompanied by instructions or primes to think about progress toward goals (Geng et al., 2016; Sussewind & Hoelzl, 2014), or (c) involved simply performing or recalling a behavior without additional instructions to reflect on these actions in a particular way (e.g., Jordan et al., 2011). Identity-focused manipulations were those that involved instructions or priming conditions aimed to induce participants to (a) think about the behavior abstractly (e.g., focus on moral values or rules or temporally distant behavior; Conway & Peetz, 2012; Cornelissen et al., 2013), (b) think about

long-term traits (e.g., write about their traits; Sachdeva et al., 2009), (c) think about commitment to, or how to achieve long-term goals (e.g., Susewind & Hoelzl, 2014). This category also included manipulations that involved providing participants with feedback that ostensibly reflected their identity or character, for example, telling them that they are “helpful” or “environmental” people (Cornelissen et al., 2007; Guadagno et al., 2001). Where manipulations included elements of both behavior- and identity-focused manipulations they were left unclassified and

omitted from the analysis of this moderator. For example, Mann and Kawakami (2011) provided false feedback that participants were progressing toward a goal (behavior-focused) of becoming less prejudiced people (identity-focused). Target Behavior Type. Studies were coded according to whether the target behavior involved the performance of actual behavior (e.g., donating real money to charity) or hypothetical behavior (e.g., responses to hypothetical dilemmas). Outcomes that involved asking the participant to agree to perform a behavior in the future (e.g., sign up to volunteer; Kristofferson et al., 2014) were coded as real behavior when they were presented to participants as genuine requests for future action rather than as hypothetical requests (i.e., participants believed they would actually engage in the behavior in the future; Goldman et al., 1982) and as hypothetical when they were framed as general intentions (e.g., future prosocial intentions; Jordan et al., 2011). Area of Literature. Studies were coded according to whether they came from the prosocial-FITD literature or whether they were published outside of this literature (non-FITD). Categorization was based on the authors’ description of their research (i.e., whether they identified it as an FITD study or not). Audience for Manipulation and Target Behavior. Studies were coded as having separate audiences for the manipulation and target behavior if (a) participants were explicitly told that the manipulation and target behavior were part of separate, unrelated studies or (b) the manipulation and target behavior were presented by different, ostensibly unrelated people. Studies in which no attempt to create the impression of separate audiences was made were coded as having one audience. Studies in which there was insufficient

FERGUSON, KAUFMANN, BROWN, AND DE LA PIEDAD GARCIA

information to determine whether an attempt to portray separate audiences had been made or where the number of audiences was unclear were omitted from this moderation analyses. Year. Year was coded as per the official publication date on journal websites. For unpublished studies, year was coded as the year data were collected or the year unpublished articles were prepared.

Meta-Analytic Procedure Effect Size Data and Calculations for Individual Studies. The effect size metric used in the current analyses is Hedges’ g (a standardized mean difference that has been corrected for sampling bias; Hedges, 1981), reported with 95% confidence intervals. All effect sizes were coded such that positive values indicated a moral licensing effect (i.e., an increase in subsequent negative behavior after a manipulation, relative to the control) and negative values indicated a positive consistency effect (i.e., a decrease in subsequent negative behavior after a manipulation, relative to the control). We interpret the magnitude effects by comparing them to percentiles of effects typically reported in social psychology (Lovakov & Agadullina, 2021), as SMB studies are typically published in social psychology or adjacent fields (e.g., organizational behavior). For continuous target behaviors (e.g., amount of cheating; Mazar & Zhong, 2010), effect sizes were calculated from cell sizes, means and variance (SDs or SEs) wherever possible, or inferential statistics (e.g., t tests) when descriptive statistics were not available. For dichotomous target behaviors (e.g., agree to volunteer or not; Kristofferson et al., 2014), we extracted cell sizes and frequencies where available, or chi-square statistics where frequencies were unavailable. These data were entered into the program Comprehensive Meta-Analysis (Borenstein et al., 2014) to calculate Hedges’ g values and effect size variance. Where insufficient statistical data were reported in study records, we contacted study authors to request missing information. If no response was received, then data were estimated, when possible (e.g., dividing the overall sample size by the number of conditions), or the study was excluded if estimation was not possible. We include notes indicating which effect sizes required estimation in the additional online material (Table S3: <https://osf.io/yb589>). Point Estimate Calculations. The individual effect sizes extracted from included studies were aggregated into point estimates (i.e., average effect sizes) using a random-effects model with robust variance estimation (RVE; Hedges et al., 2010). As with individual effect sizes, the point estimate used was Hedges’ g with 95% confidence intervals. Effect sizes were coded such that a positive point estimate indicates moral licensing (i.e., an increase in negative behavior), whereas a negative point estimate indicates positive consistency (i.e., a decrease in negative behavior). Heterogeneity between effect sizes was quantified with τ , which represents the estimated standard deviation of the aggregated effect size estimate (Borenstein et al., 2009). We selected RVE methods to combine effect sizes to account for two types of

dependency present in the data: correlated effects and hierarchical relationships between effect sizes and studies. RVE accounts for these kinds of dependency by adjusting effect size weights, even when the degree and structure of covariance between effect sizes are unknown (Hedges et al., 2010). First, we used a correlated effects model (CEM; Fisher & Tipton, 2015) to account

for dependency due to correlated effects. A small but notable number of effect sizes ($k = 19$) were calculated from studies that included multiple moral manipulation conditions, but a single control group (e.g., Experimental Group 1 vs. Control, Experimental Group 2 vs. Control). In these cases, effect sizes were calculated for each moral manipulation condition in the study. As these effect sizes are calculated using the same control group, they are correlated and thus nonindependent. CEM models account for this dependency by calculating weights based on sample clusters. Effect sizes derived from shared control groups are grouped into the same sample clusters, and effect sizes derived from independent samples are counted as individual clusters. Second, we used a hierarchical effects model (HEM; Fisher & Tipton, 2015) to account for dependency due to hierarchical relationships between effect sizes and studies. There were a large number of effect sizes calculated from independent groups of participants that were nested within studies. These studies included multiple moral manipulation conditions and multiple control groups (e.g., Experimental Group 1 vs. Control Group 1; Experimental Group 2 vs. Control Group 2). In cases like this, even though effect sizes are based on the different groups of participants, there may still be dependency between them as they emerged from the same studies. HEM estimates account for this kind of dependency by calculating weights based on study clusters. Effect sizes derived from the same studies are grouped into the same clusters. The results of both CEM and HEM RVE models are presented throughout the results. RVE analyses were conducted using the *robumeta* package in R (Fisher & Tipton, 2015). Outliers. We ran influence diagnostics using the *metafor* package in R to identify outliers and quantify their influence on the data before combining effect sizes (Viechtbauer, 2015). Cases with studentized residuals $> \pm 1.96$ were identified as outliers. Outliers with Difference in Fits (DFFITs) values $> \pm 0.20$ (i.e., $3 \times k - p$ where p is equal to the number of estimated coefficients and k is equal to the number of observations) were considered influential (Viechtbauer & Cheung, 2010). Influential cases were removed prior to analyses. Moderation Analyses. Moderators were coded categorically (see descriptions above) and dummy-coded as 0 and 1 (see Table 3). Manipulation behavior type consisted of three levels (hypothetical, recalled, or actual behavior), so two dummy-coded comparisons were created using the hypothetical manipulations as the reference category (coded as 0). Year was entered as a continuous moderator. We tested whether effect sizes differed systematically by moderator group using metaregression with RVE using both CEM and HEM weights. We began by testing each moderator category separately. We then entered significant moderators simultaneously into a metaregression to identify whether each moderator uniquely accounted for differences in effect sizes, or whether the results of the initial simple metaregressions were partially driven by overlap between moderator categories. File-Drawer Bias. File-drawer bias refers to the issue of missing data from unreported studies that are typically more likely to find smaller effects and nonsignificant results (Rosenthal, 1979). Omitting these studies from meta-analyses can result in overestimation of the reliability and magnitude of effect sizes. We did receive some reports of unpublished results from authors in the field (~26% of effect sizes included in Meta-Analysis 1 were unpublished. The unpublished

SmC

k

0.04 -0.09

0.11 -0.44 [-0.04, 0.11] [-0.31, 0.14]

[0.05, 0.17] [-0.59, -0.29]

[-0.006, 0.16] [-0.09, 0.08]

[-0.17, 0.04] [-0.03, 0.11]

-0.06 0.04 0.07 0.01

[0.08, 0.26] [-0.10, 0.11] [-0.27, -0.05]

[0.04, 0.30] [-0.15, 0.03]

[0.16, 0.44] [-0.14, 0.004]

95% CI

0.31 0.47

0.27 0.35

0.28 0.34

0.27 0.32

0.27 0.25 0.37

0.31 0.33

0.23 0.32

τ QB [95% CI]

0.01 [0.001, 0.01]

-0.09 [-0.32, 0.11]

-0.54 [-0.71, -0.38]

-0.08 [-0.20, 0.04]

0.10 [-0.04, 0.23]

-0.16 [-0.29, -0.02] -0.32 [-0.46, -0.18]

-0.23 [-0.39, -0.07]

-0.32 [-0.44, -0.21]

Correlated effects RVE model

0.17 -0.01 -0.16

0.17 -0.06

0.25 -0.07

g

StC

0.03 -0.09

0.11 -0.43

0.07 -0.02

-0.06 0.04

0.17 -0.003 -0.17

0.16 -0.07

0.24 -0.07

g

[-0.04, 0.10] [-0.32, 0.14]

[0.05, 0.17] [-0.54, -0.33]

[-0.01, 0.15] [-0.11, 0.07]
 [-0.17, 0.04] [-0.03, 0.11]
 [0.08, 0.26] [-0.11, 0.10] [-0.28, -0.07]
 [0.04, 0.29] [-0.16, 0.02]
 [0.15, 0.33] [-0.15, -0.003]

95% CI

0.30 0.30

0.27 0.40

0.28 0.33

0.27 0.31

0.27 0.25 0.32

0.19 0.32

0.23 0.32

T

0.01 [0.001, 0.02]

-0.10 [-0.34, 0.14]

-0.53 [-0.64, -0.42]

-0.09 [-0.21, 0.03]

0.11 [-0.03 0.24]

-0.17 [-0.31, -0.04] -0.33 [-0.47, -0.20]

-0.23 [-0.38, -0.08]

-0.32 [-0.43, -0.20]

QB [95% CI]

Hierarchical weights RVE model

Note. Significant effect sizes and regression coefficients are bolded. RVE = robust variance estimation; k = total effect sizes; SmC = sample clusters; g = Hedges' g average effect size estimates for studies within moderator subgroup (positive values represent a moral licensing effect; negative values represent a positive consistency effect); CI = confidence interval; QB = regression coefficients for tests of moderators; StC = study clusters; FITD = foot-in-the-door.

Target behavior valence Negative (0) Positive (1) Domain-consistency Inconsistent (0) Consistent (1) Manipulation behavior type Contrast 1: Hypothetical (0) versus recall (1) Contrast 2: Hypothetical (0) versus behavior (1) Hypothetical Recall Behavioral Manipulation focus Identity-focused (0) Behavior-focused (1) Target behavior type Hypothetical (0) Behavior (1) Area of literature Non-FITD (0) Prosocial-FITD (1) Audience One audience (0) Two audiences (1) Year

Moderator

Table 3 Meta-Analysis 1: Average Effect Size and Heterogeneity Estimates for Full Sample and Moderator Categories

REVIEW OF SEQUENTIAL MORAL BEHAVIOR

studies came from a small number of researchers and were exclusively for non-FITD studies. As such, it is likely these represent only a portion of unreported SMB studies that have been conducted. Thus, we were interested in assessing whether our analyses were impacted by file-drawer bias. We emphasize here that we were not interested in assessing if publication bias was present in the published literature, which would involve only analyzing published results. Instead, we were interested in evaluating if our analysis, which incorporated both published and unpublished effects, may be biased because there are still unreported studies that we were unable to locate and include. To that end, we conducted bias analyses on the full data set, including both published and unpublished effects. We used contour-enhanced funnel plots to examine if bias may be present in the data (e.g., Palmer et al., 2008). These graphs plot effect size against standard errors and then overlay contours that represent regions of conventional statistical significance. Unreported studies are typically more likely to be nonsignificant, so if studies appear to be absent from regions of nonsignificance, this suggests that significant results are overrepresented in the effect sizes included in the analyses, and nonsignificant results (that would be expected even if an effect is real) are systematically missing. Contour-enhanced plots were produced using the *metafor* package in R (Viechtbauer, 2015). Next, we conducted precision-effect test and precision-effect estimate with standard errors (PET-PEESE; Stanley & Doucouliagos, 2014) and a three-parameter selection model (3-PSM; Hedges & Vevea, 1996) to statistically identify if estimates are likely to be impacted by file-drawer bias and calculated revised effect size estimates that attempt to correct for this bias. PET-PEESE uses a weighted regression model (linear for PET, quadratic for PEESE) that predicts effect sizes from their standard errors. If the slope of the regression model is significant, this indicates the data are likely impacted by file-drawer bias. If this is the case, the model intercept provides an effect size corrected for file-drawer bias. When selecting whether to use PET or PEESE estimates, we adhered to conditional logic recommended by Stanley and Doucouliagos (2014): When PET produced a nonsignificant effect size estimate, we retained these estimates; where PET produced a significant effect size estimate, we used PEESE estimates instead. PET-PEESE was conducted using the *metafor* package in R (Viechtbauer, 2015). We used a 3-PSM (Hedges & Vevea, 1996) using the *weightr* package in R (Coburn & Vevea, 2017). 3-PSM models file-drawer bias by estimating the likelihood of significant results being present (selection parameter) and then estimating the mean effect size and heterogeneity of effect sizes (data parameters) when file-drawer bias is present. If the fit of the model that assumes file-drawer bias is significantly better than the fit of a model that assumes no bias, then this is taken as evidence of bias and the modeled effect size estimate is interpreted as a corrected “true” effect size estimate.