Psychological Methods
2015, Vol. 20, No. 3, 310–330

1082-989X/15/$12.00 http://dx.doi.org/10.1037/met0000046

Publication Bias as a Function of Study Characteristics

Kathleen M. Coburn and Jack L. Vevea
The University of California, Merced

Researchers frequently conceptualize publication bias as a bias against publishing nonsignificant results. However, other factors beyond significance levels can contribute to publication bias. Some of these factors include study characteristics, such as the source of funding for the research project, whether the project was single center or multicenter, and prevailing theories at the time of publication. This article examines the relationship between publication bias and 2 study characteristics by breaking down 2 meta-analytic data sets into levels of the relevant study characteristic and assessing publication bias in each level with funnel plots, trim and fill (Duval & Tweedie, 2000a, 2000b), Egger's linear regression (Egger, Smith, Schneider, & Minder, 1997), cumulative meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009), and the Vevea and Hedges (1995) weight-function model. Using the Vevea and Hedges model, we conducted likelihood ratio tests to determine whether information was lost if only 1 pattern of selection was estimated. Results indicate that publication bias can differ over levels of study characteristics, and that developing a model to accommodate this relationship could be advantageous.

Keywords: publication bias, funnel plot, trim and fill, weight-function model, cumulative meta-analysis

Supplemental materials: http://dx.doi.org/10.1037/met0000046.supp

Traditionally, researchers define publication bias as the bias in effect size estimates that occurs because a study's likelihood of publication depends on its results' direction or statistical significance (Rothstein, Sutton, & Borenstein, 2005). Most methods for assessing or compensating for publication bias, including trim and fill (Duval & Tweedie, 2000a, 2000b), Begg and Mazumdar's (1994) rank correlation test, Egger's linear regression (Egger et al., 1997), weight-function methods (e.g., Vevea & Hedges, 1995), and fail-safe *N* (Rosenthal, 1979) have all assumed that publication bias is a function only of sample size or statistical significance. However, there is a substantial body of evidence demonstrating that publication bias can be due to other factors, including study characteristics other than significance. Cooper, DeNeve, and Charlton (1997) interviewed 33 investigators about the fate of 159 institutional review board approved studies and noted, "This study confirms the existence of filters in the research process other than bias against the null hypothesis that act to select studies out of the publication process" (p. 451). These "filters" may include the source of funding for the research, the time when the research is conducted and corresponding social preferences, whether a result is the primary or secondary focus of a study, whether research is conducted at a single center or at multiple centers, and even the gender of the principal investigator.

Industry funding of studies, or cases in which primary investigators have financial interests in the outcome of analyses, can have a strong influence on whether research is published, regardless of the significance of its results. Roseman et al. (2011) note that "conflicts of interest may influence the framing of research questions, study design, data analysis, interpretations of findings, whether to publish results, and what results are reported" (p. 1008). Some research topics[1] in fields such as industrial and organizational (I-O) psychology can be strongly affected by funding-related bias, such as when test vendors choose not to release any information about their measure, particularly effect size information (McDaniel, Rothstein, & Whetzel, 2006). Publication bias in I-O psychology can affect the efficacy of human resource management practices and the effectiveness of organizations (Kepes, Banks, McDaniel, & Whetzel, 2012).

In medical research, industry-sponsorship of clinical trials tends to yield proindustry conclusions, and industry-sponsored trials are more likely to use nonactive or placebo controls than nonindustry trials, which increases the likelihood of positive results (Bekelman, Li, & Gross, 2003). Publication bias in this fashion can lead to an inaccurate public impression of the efficacy of a drug, to misuse of drugs, or to false impressions of a drug's negative side effects. One meta-analysis coded 106 review articles on the health effects of passive smoking for variables including peer review status, author affiliation (with or without tobacco companies), article topic (the health issue at hand), and year of publication. Of all variables coded, only affiliation with the tobacco industry was a significant predictor of finding passive smoking not harmful ($p < .05$). Only

Kathleen M. Coburn and Jack L. Vevea, Psychological Sciences, School of Social Sciences, Humanities and Arts, The University of California, Merced.

Thanks to Rose Scott, The University of California, Merced, for her comments and support.

Correspondence concerning this article should be addressed to Kathleen M. Coburn, Psychological Sciences, School of Social Sciences, Humanities and Arts, The University of California, Merced, 1100 North Lake Road, Merced, CA 95343. E-mail: kcoburn@ucmerced.edu

[1] Whereas it is not uncommon to view publication bias as affecting an entire field, Banks, Kepes, and McDaniel (2015) argue that it is often more appropriate to consider whether it affects specific topics. We support that view.

6% of tobacco-affiliated authors found passive smoking harmful, compared with 87% of nonaffiliated authors (Barnes & Bero, 1998). In addition, trials sponsored by pharmaceutical companies are published more prominently (Doucet & Sismondo, 2008). Source of funding, therefore, clearly can play a powerful role in the outcome of studies and their likelihood of publication across disciplines.

Social preferences at the time when research is conducted can also affect the likelihood of publication, regardless of the direction of results. Kepes, Banks, and Oh (2014) write, "The assumptions underlying publication bias suggest that results are suppressed if they are contrary to trends of past research or beliefs held by researchers, editors, or reviewers" (p. 196). Begg and Berlin (1988) noted, "In general, the consequences of publication bias may be the most deleterious in circumstances where there is no real effect, but where social preferences are leading to selective publication" (p. 442). Glass, McGaw, and Smith (1981) discussed a meta-analysis conducted by Smith (1980) assessing whether sex bias was present among psychological counselors—that is, whether counselors held concepts about the nature of women that reflected societal stereotypes and consequently pressured their female clients into fulfilling those stereotypes. This meta-analysis uncovered a surprising effect: The journal articles demonstrated a bias against women, whereas the unpublished dissertations demonstrated a bias in the opposite direction, favoring women. Taken as a whole, the average effect was near zero. This did not seem to be an artifact of quality, because the dissertations received higher quality ratings on average. From this, Glass et al. concluded, "The studies most likely to be submitted and/or accepted for publication are those that demonstrate sex bias against women [supporting the pervasive belief at the time], regardless of their quality" (p. 51).

Social preferences may have a particularly strong impact in such areas as race and gender differences. For example, research on the magnitude of Black/White mean differences in job performance often underestimates the effect, possibly because authors are unwilling to report mean racial differences (McDaniel, McKay, & Rothstein, 2006). Published journal articles typically report smaller mean racial differences than unpublished technical reports (McKay & McDaniel, 2006). In 1994, Herrnstein and Murray published *The Bell Curve*, a book that took a controversial stance on the relationship between race and intelligence, arguing that genetics play a role in racial IQ differences. The publication of this book is another example of the influence of shifting social preferences on publication bias. Prior to the release of *The Bell Curve*, racial differences in intelligence were a subject rarely discussed and seldom addressed in publication. *The Bell Curve* opened the door to a wealth of academic discussion and debate, and to a corresponding wealth of new publications. Of this changing trend, Heckman (1995) wrote,

> Within academia, it has been folly for scholars in pursuit of peer-reviewed publication and peer-reviewed funding to critically examine these issues. By raising these and other forbidden topics in a best-selling book, Herrnstein and Murray perform a valuable service for the academic community. They have written a bull's-eye for academics to fire at and have written it in such a way that they are certain to draw fire. (p. 1092)

(One could argue that this "valuable service" extends beyond the academic community.) As for gender differences, Kepes, Banks, and Oh (2014) discuss an example (Eagly, Johannesen-Schmidt, & van Engen, 2003) in which selective publication may artificially create the impression that women display more transformational leadership behaviors than men.

Authors are not alone in this type of bias: Journal reviewers also may treat harshly research findings that challenge previously held beliefs. Although reviewers are not likely to cite personal disagreement as a reason for rejection and may not even be aware of their prejudice, experimental studies (Abramowitz, Gomes, & Abramowitz, 1975; Goodstein & Brazis, 1970; Mahoney, 1977) demonstrate that this kind of bias does exist (Armstrong, 1996). This is often referred to as conformity publication bias, in which confirmatory studies are published and studies that contradict currently held beliefs, regardless of significance, are not. The reverse is possible if new studies contradict a currently held belief and are published because they are newsworthy (Felson, 1992). The investigators' own impression of the importance of their results can influence the likelihood of publication as well; Easterbrook, Gopalan, Berlin, and Matthews (1991) found that "studies with a high importance rating by the investigator were significantly more likely to be published than those with a low rating" (p. 870). In short, "the point is that publication bias may be a function not just of statistical significance but of the ebb and flow of editorial and consensus opinion" (Felson, 1992, p. 887).

Other factors such as use of randomization and the number of sites involved in a study can influence the likelihood of publication. Use of randomization significantly predicts publication such that observational, laboratory-based experimental studies and nonrandomized trials have a greater risk of publication bias than randomized clinical trials (Begg & Berlin, 1988; Berlin, Begg, & Louis, 1989; Dickersin & Min, 1993; Easterbrook et al., 1991). Begg and Berlin (1989) suggest,

> Since randomized trials generally involve a greater commitment of time, authors may be more inclined to follow through to publication. Conversely, historic comparisons can be assembled from a database with relative ease and are therefore easily discarded if the results are uninteresting. (p. 109)

Multicenter trials are somewhat less prone to bias than studies conducted in a single institution, possibly due to increased pressure in the cooperative groups for publication regardless of the results of the trial (Begg & Berlin, 1988; Berlin et al., 1989).

There is, then, a body of research that provides evidence that publication bias may differ across levels of study characteristics. The purpose of this project is to select data sets from two previously published meta-analyses and, using five different methods of assessing publication bias, to demonstrate empirically that study characteristics can affect studies' likelihood of publication.

Method

We aimed to locate two meta-analyses as examples, one in which differing patterns of bias were an issue and one in which they were not, to illustrate that this situation does exist in the literature and to demonstrate techniques for addressing it. With this in mind, we briefly reviewed the past 45 volumes (published from 1983 to 2013) of *Psychological Bulletin*. We chose *Psychological Bulletin* as the subject of this search because it is a journal dedicated to the publication of research syntheses and meta-

analyses. We want to emphasize that our search was not systematic and did not stop once we identified two studies; rather, we browsed the issues for two meta-analyses with suspect study characteristics to examine. We searched only for meta-analyses that published their effect sizes so that we could easily reanalyze the data, and we considered only meta-analyses with a hundred or more total effect sizes because assessing publication bias with the Vevea and Hedges (1995) weight-function model requires a relatively large sample of effects to work well. In addition, because each level of the study characteristic would also be assessed with the weight-function model, there needed to be enough effect sizes in each level. Hence, candidate meta-analyses needed to be sufficiently large, with a moderator variable that might affect patterns of publication bias, and with at least 50 to 60 effect sizes per level of the moderator.

We read the abstracts of all articles that mentioned either *synthesis* or *meta-analysis* in the title and set aside any meta-analyses with sufficiently large data sets for further investigation. We ultimately chose two meta-analyses examining gender differences; both of these articles discussed and coded for a study characteristic that might be related to publication bias. Once we chose these articles, we extracted both data sets and divided each into two groups corresponding to levels of the relevant study characteristics. Then we assessed publication bias three times per data set—once per group and once for the overall data set—using five different methods, discussed below.[2] We conducted all the analyses using the open-source statistical software R version 3.1.2 (R Core Team, 2013). Below are descriptions of the data sets we identified and their relevant study characteristics, followed by an overview of the five methods used. These data sets and code to run the corresponding analyses are included as online supplemental materials.

It is important to keep in mind that both data sets are heterogeneous. Therefore, several methods that do not work as well in the presence of heterogeneity, namely funnel plots, trim and fill, and Egger's regression, may be inaccurate. Weight-function models are capable of accommodating heterogeneity and can outperform such methods as trim and fill and other funnel plot-based assessments in simulations if heterogeneity is present (Terrin, Schmid, Lau, & Olkin, 2003). Cumulative meta-analysis, similar to weight-function models, is also less affected by heterogeneity (Kepes, Banks, & Oh, 2014). Ultimately, the application of all five methods should be considered a sensitivity analysis, or an assessment of whether the meta-analytic models estimated are robust to the effects of publication bias.

Data Sets

Hyde and Linn (1988): "Gender Differences in Verbal Ability: A Meta-Analysis." This meta-analysis assesses the popular belief that gender differences in verbal ability favoring females are a well-established psychological finding (Hyde & Linn, 1988). The authors mention knowing little about the nature of this gender difference beyond the global statement that "females have superior verbal ability" (p. 54). They estimate a fixed-effects meta-analytic model and calculate a weighted average of $d = 0.11$, indicating that women outperform men on verbal tasks by approximately one-tenth of a standard deviation. The authors conclude that this effect is negligible, and that gender differences no longer exist. Because the effect sizes were significantly heterogeneous, however ($Q = 2196.08$, $p < .05$), the authors conduct moderator analyses and find that year of publication is a significant predictor of effect size. Studies published in 1973 or earlier had an average effect size of $d = 0.23$; in contrast, studies published in 1974 or later had an average effect size of $d = 0.10$. (Our analyses did not replicate these mean differences, which is likely due to the fact that the authors computed weighted averages rather than conducting random-effects meta-analyses.)

Hyde and Linn (1988) mention two potential interpretations of this decline in gender differences. Their first theory is that it may be due to the increasing flexibility of gender roles in recent years, such that boys are engaging in more activities that are stereotypically attributed to girls or vice versa, and their second is that it may be the result of researchers' changing publication practices. The authors note that, in 1974, Maccoby and Jacklin published a book containing a review of 85 studies on gender differences in verbal ability that pointed out researchers' tendencies not to publish nonsignificant studies of gender differences. Maccoby and Jacklin wrote, "There are instances in which there has been direct pressure to keep findings out of the published literature when they do not agree with the accepted view of some process or relationship" (p. 4). It is possible that the publication of this book encouraged researchers to report their null or nonsignificant findings, resulting in an increasing number of small effect sizes uncovered during literature searches, and subsequently in a smaller average effect size post-1974. Here, year of publication appears to be a study characteristic that might influence publication bias.

Kling, Hyde, Showers, and Buswell (1999): "Gender Differences in Self-Esteem: A Meta-Analysis." This data set comes from a meta-analysis conducted by Kling et al. (1999) of 216 effect sizes from studies that analyzed gender differences in self-esteem. The authors find a weighted mean effect size of $d = 0.21$ (95% confidence interval from 0.19 to 0.22), indicating that males on average have greater self-esteem. Homogeneity analyses are significant, indicating that the data are heterogeneous and leading to the analysis of moderators ($Q = 626.61$, $p < .05$). One of the moderating variables the authors examine is whether the gender comparison coded from each article was the primary or secondary focus of the study. They theorize that, because gender effect sizes from articles focused on a topic other than gender are peripheral to the article and are presumably reported incidentally, they are less likely to be subject to publication bias (Kling et al., 1999). The authors code 99 effect sizes from gender-focused articles and 97 effect sizes from non-gender-focused articles, a total of 196 effect sizes (20 are excluded for not meeting the criteria of either category). The gender-focused effect sizes yield a mean of $d = 0.24$, and the remaining effect sizes yield a smaller mean of $d = 0.16$. The difference between these groups is statistically significant. In this case, article focus appears to be a relevant study characteristic.

Assessments

1. Funnel plot. A funnel plot (Light & Pillemer, 1984) is a scatterplot of effect sizes against some measure of sample size or

[2] Although fail-safe *N* is a common method of assessing publication bias (Kepes et al., 2012), we did not include it here due to its documented flaws (Becker, 2005; McDaniel et al., 2006).

precision. It is termed a "funnel plot" because the precision in estimation of the effect increases as the sample size of the study increases. This will result in the appearance of a symmetrical inverted funnel about the population mean effect in the absence of publication bias if the data are homogenous—that is, if all studies estimate the same underlying effect (Sterne, Becker, & Egger, 2005).

Researchers visually assess funnel plots for symmetry in order to determine areas of asymmetry, or reduced density, in which effect sizes may be unobserved due to publication bias. Traditionally, these graphs plot a measure of study size on the vertical axis. Although this measure was originally precision, studies have demonstrated that researchers should generally use standard error or its square (i.e., the inverse of precision) instead (Sterne & Egger, 2001). If there is systematic heterogeneity due to moderator variables, the funnel plot may appear asymmetric even though publication bias is not actually present. For a demonstration of this phenomenon, see Figure 1. The first funnel plot shows signs of asymmetry; there appear to be missing effect sizes in the lower right-hand corner of the plot. The second funnel plot reveals that this asymmetry is due to a moderator. The filled-in circles were simulated with a true mean of $d = 0.70$, and the hollow circles were simulated with a true mean of $d = 0.20$ and larger sample sizes. Considered separately, neither the filled-in nor hollow circles show any sign of asymmetry.

We plotted effect sizes (here, in the form of standardized mean differences) against their corresponding standard errors; however, we chose to plot effect size magnitude on the vertical axis and standard error on the horizontal axis. This plotting procedure results in a funnel shape that is horizontal rather than vertical, a decision made in order to allow the effect sizes to spread out and to ease assessment of symmetry (Sterne & Egger, 2001). We added contour lines to our plots to create what are known as contour-enhanced funnel plots. Contour-enhanced funnel plots are drawn with lines distinguishing effect sizes with different levels of statistical significance to help determine what degree of asymmetry is due to publication bias rather than unidentified moderators of effect size (Peters, Sutton, Jones, Abrams, & Rushton, 2008). We have added three sets of contour lines to each funnel plot, distinguishing effects with $p < .01$, $.01 < p < .05$, and $.05 < p < .10$.

We constructed three funnel plots for each meta-analysis in R using the *plot* function—one for the complete data set and two more for the effect sizes in each level of the study characteristic. We assessed symmetry in all three plots and compared patterns of asymmetry across plots.

2. Trim and fill. Duval and Tweedie (2000a, 2000b) created a nonparametric method of estimating the number of studies unobserved, or missing, from a data set due to publication bias. This method, called *trim and fill*, relies on the idea that a funnel plot will feature a symmetric distribution of effect sizes around the population mean effect in the absence of publication bias, and assumes that those effect sizes that go unpublished will be the most extreme effect sizes (Jennions & Møller, 2002). It uses an iterative process to determine the symmetric portion of the funnel plot, to trim or remove the asymmetric effect sizes, and then to use one of three estimators (referred to as L_0, R_0, and Q_0) to generate the missing studies necessary to achieve symmetry (Jennions & Møller, 2002). Trim and fill then produces both an unadjusted mean estimate and an adjusted mean estimate for the data set (incorporating the filled-in studies, which are assigned the same sampling variance as their trimmed counterparts). The number of filled-in effect sizes represents the approximate number of effect sizes theoretically suppressed by publication bias. Trim and fill generally assumes homogeneity of effect sizes, although a modification exists to incorporate moderators in the context of metaregression (Weinhandl & Duval, 2012). This proposed algorithm is new and the R function to implement it, *trim.fill.regression*, can accommodate only one moderator (Duval & Weinhandl, 2011). Because we were able to obtain the *trim.fill.regression* function (Duval & Weinhandl, 2011), we assessed publication bias with it as well, as described below.

There are at least two other functions that estimate trim and fill in R. We used the *trimfill* function of the R package metafor (Viechtbauer, 2010) because it permits users to specify on which side of the funnel plot effect sizes are missing. We conducted trim-and-fill sensitivity analyses on the full data sets and on both groups individually, always iterating from and estimating a fixed-effects model. The original formulation of trim and fill iterated from a random-effects model, but later simulations indicated the superior performance of trim and fill when iterating from a fixed-effects model (Weinhandl & Duval, 2012). Iterating from a random-effects model can incorrectly adjust asymmetry and give samples with less precision too much weight (Kepes, Banks, & Oh, 2014). We performed each analysis four times, twice using the L_0 estimator (once for censorship on each side of the plot) and twice using the R_0 estimator (similarly), resulting in 12 trim-and-fill analyses per data set. Both estimators were used because simulations have shown R_0 to be more accurate over a larger range of k (where k is the number of effect sizes), whereas L_0 performs slightly better if k is smaller (Duval & Tweedie, 2000b). We also used both estimators because Duval and Tweedie recommend using both before making a judgment on the actual number that might be suppressed. We did not use the third estimator, Q_0, because L_0 is a simple linearization of Q_0 (Duval, 2005).[3]

We also used the *trim.fill.regression* function to conduct sensitivity analyses, again iterating from a fixed-effects model in all cases. This function requires that users specify a moderator variable; here we created a dummy-coded variable representing group membership. The function then iterates from a fixed- or random-effects model and produces a matrix of results, which contains the results of a set of fixed- and random-effects models estimated on both the observed data and the imputed data (Weinhandl & Duval, 2012). It also provides information on the number of effect sizes added for each level of the moderator. Theoretically, if trim and fill with a metaregression algorithm adds more studies to one level than to another, this can provide some evidence that the pattern of publication bias differs across levels. Early applications of this function in simulation assessed its ability to accommodate moderators of effect size (Weinhandl & Duval, 2012), so little is known about its ability to differentiate between patterns of bias, and future simulations in this area might be useful. Here we present the results of trim and fill with metaregression using both the L_0 and R_0 estimators and both maximum likelihood (MLE) and

[3] A more in-depth explanation of the performance of these estimators is not necessary to understand the following analyses; however, one can be found in Duval and Tweedie (2000b).

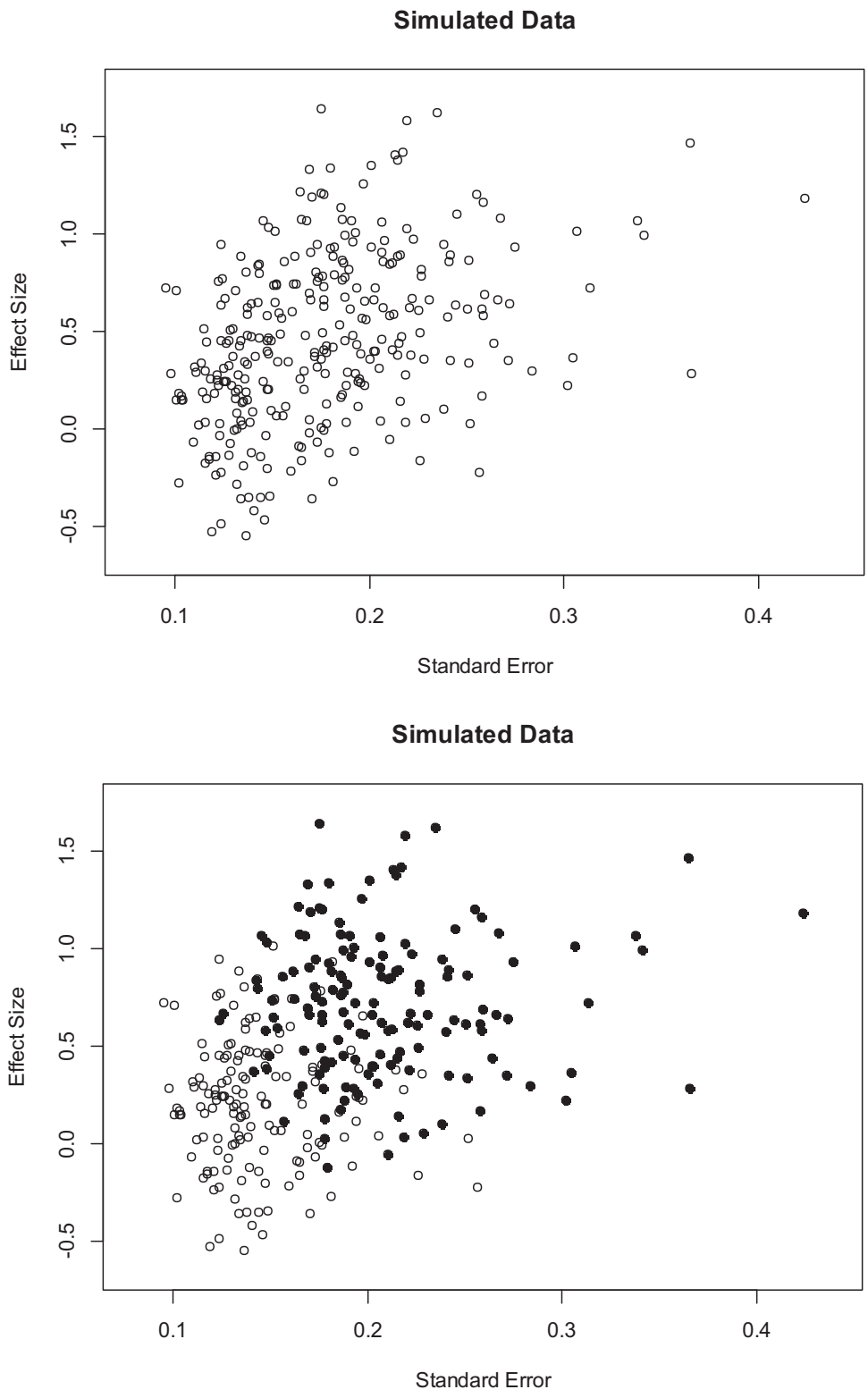

Figure 1. Funnel plots demonstrating the influence of systematic heterogeneity. Data simulated using a random-effects model with a variance component (τ^2) of 0.03. Filled-in circles have a true mean of $d = 0.70$; hollow circles have a true mean of $d = 0.20$ and larger sample sizes.

method of moments (MOM). These results include an estimate of the adjusted intercept and slope and the number of effect sizes imputed to each level.

3. Egger's linear regression. Egger et al. (1997) introduced a linear regression approach for examining associations between study size (or sample size) and estimated effect sizes. This approach is equivalent to a weighted regression of effect size on standard error, with weights that are inversely proportional to the variances of the effect sizes (Sterne, Egger, & Smith, 2001). This method may be understood as a statistical analogue of a funnel plot (Sterne et al., 2001), and it assumes that effect sizes are homogenous, or that all effect sizes estimate the same underlying "true" mean effect. (Indeed, all publication bias assessments that are based on the funnel plot are vulnerable to distortion if there is systematic heterogeneity.) It is found to be more sensitive than its counterpart, the rank correlation approach (Begg & Mazumdar, 1994), but its sensitivity is reduced in meta-analyses where $k < 20$ (Sterne et al., 2001). The idea is that this regression does not test directly for publication bias, but for a relationship between the observed effect sizes and standard error. If a relationship is present, this usually implies asymmetry in the plot, which can in turn indicate the presence of publication bias (Viechtbauer, 2010). However, publication bias is not the only possible cause of asymmetry, and Egger's linear regression does not incorporate predictors that represent study characteristics other than precision. An adaptation has been proposed in the context of metaregression that may be able to accommodate systematic heterogeneity via a quadratic approximation without a linear term (Stanley & Doucouliagos, 2014). Because the code to implement that method is not publicly available, though, it was not implemented here.

We used the *regtest* function of the R package metafor (Viechtbauer, 2010) to conduct Egger's linear regression three times per data set—once for the complete data set and once for each individual level of the relevant study characteristic. We chose this function because it allows users to specify whether the model used should be a random-effects model or a traditional linear model. For each analysis, we specified a traditional linear model and used standard error as a predictor.

4. Cumulative meta-analysis. Cumulative meta-analysis is another tool for visually assessing publication bias, this time without using a funnel plot. It is not strongly affected by between-studies heterogeneity (Borenstein et al., 2009; Kepes et al., 2012; McDaniel, 2009). The analyst sorts effect sizes by a characteristic of interest. For the assessment of publication bias, this is usually precision, to investigate the existence of small-study effects. The analyst adds effect sizes one at a time and recalculates the meta-analysis as each new effect size is added (Borenstein et al., 2009). This results in a situation where the most precise effect is added first, followed by the next most precise, and so on (Kepes, Banks, & Oh, 2014). Then a forest plot of these means is examined for drift, or gradual movement in the cumulative mean effect estimate (Borenstein et al., 2009). If there is positive drift (drift to the right-hand side), there is evidence that effect sizes of small magnitude from studies with small sample sizes are being suppressed, which is indicative of a small-study effect, or publication bias (Kepes, Banks, & Oh, 2014). Because this technique is relatively recent, there are no clear guidelines yet for these visual judgments (Kepes, Banks, & Oh, 2014). Kepes et al. (2012), though, do mention the utility of interpreting the drift in light of other indicators of publication bias such as trim-and-fill results, and of comparing the first mean estimate with the last to assess the severity of drift. Cumulative meta-analysis is useful in cases where the number of studies is small and tests based on funnel plots or selection models may be underpowered.

For this assessment, we sorted the effect sizes by precision. We then estimated two cumulative meta-analyses per data set using fixed-effects models, one each on the two levels of each study characteristic, and assessed the plots visually for drift. We conducted these analyses using the *metagen* and *metacum* functions of the R package meta (Schwarzer, 2015).

5. Vevea and Hedges (1995) weight-function model. There have been several endeavors to model publication bias using weighted distribution theory, in which a weight function describes the likelihood of effect sizes in a given range being observed; Iyengar and Greenhouse (1988) proposed the first. Others modified this selection model to examine what might be termed extreme cases of weighted distributions, which estimated one weight that represented the likelihood of observing significant effects relative to nonsignificant effects (Hedges, 1984; Lane & Dunlap, 1978). This work led to the creation of models that estimated weights for a series of specified discontinuities, or intervals, with boundaries where psychologically important p values, like $p = .05$ and $p = .50$, occur (Hedges, 1992; Vevea & Hedges, 1995). The weight-function model used in this project is an example of the latter, modified to accommodate an effect-size model that can incorporate predictors (Vevea & Hedges, 1995).

The Vevea and Hedges (1995) weight-function model first estimates an unadjusted random-effects model that can incorporate predictors of effect size, where the observed effect sizes are assumed to be normally distributed as a linear function of predictors, with variance equal to their (approximately known) sampling variance plus an estimated variance component. This mixed-effects model can be reduced to a mean-only random-effects model in the absence of predictors by estimating a model that includes only an intercept and a variance component. The random- and mixed-effects models are no different from the traditional meta-analytic models. The new element of the Vevea and Hedges (1995) model is an adjusted model that can estimate not only the original mean model, either random- or mixed-effects, but also a series of weights for some prespecified p value intervals of interest. This will produce mean, variance component, and covariate estimates that have been adjusted for publication bias, as well as weights that reflect the likelihood of observing effect sizes in each designated interval. During interpretation, however, it is important to keep in mind that the weight for each estimated interval is relative to the first interval, which is always fixed to one so that the model is identified. It is also crucial to note that the model uses p value intervals corresponding to one-tailed p values. This does not imply that the model assumes selection is always one-tailed; rather, it allows flexibility in the selection function, which need not be symmetric for effects in opposite directions. A two-tailed p value of .05 is therefore represented as $p < .025$ or $p > .975$. If a funnel plot provides evidence that one-tailed of the distribution is mostly absent, we can attempt to capture that selection function by focusing in the desired direction.

An example may help to clarify. One might estimate a mean-only random-effects model with two intervals, where the discontinuities are at $p = .05$ and $p = 1.00$, distinguishing between

significance and nonsignificance. In that case, the weight for the first interval ($.00 < p < .05$) is automatically fixed to one prior to estimation. The model output then consists of three estimated parameters—the mean and variance component for the random-effects model, adjusted for publication bias, and the weight for the second interval ($.05 < p < 1.00$). An estimated weight of 0.50 for the second interval ($.05 < p < 1.00$) would indicate that nonsignificant effect sizes were half as likely to be observed as significant ones. To expand, imagine that the same model is estimated with four intervals, with discontinuities at $p = .05$, $p = .20$, $p = .50$, and $p = 1.00$. The weight for the first interval, $p < .05$, is again fixed to one. This model would produce five estimated parameters—the adjusted mean and variance component and the last three weights. Estimated weights of 0.75, 0.50, and 0.25, respectively, would indicate that effect sizes with p values between .05 and .20 were three fourths as likely to be observed as significant ones, those with p values between .20 and .50 half as likely, and those with p values between .50 and 1.00 a quarter as likely. These examples can be extended to mixed-effects models by the addition of extra parameters for covariates.

For the purposes of this project, we refer to the Vevea and Hedges (1995) model as "the weight-function model."[4] We estimated the weight-function model three times per data set, once for the full data set and once for each individual group. We specified four intervals with discontinuities at $p = .05$, $p = .20$, $p = .50$, and $p = 1.00$. We included discontinuities at $p = .50$ because, in the context of one-tailed p values, that boundary represents the point at which many effect sizes (e.g., correlations, standardized mean differences, log odds ratios) become negative and because both meta-analytic data sets featured primarily positive effect sizes, with a sharp shift in funnel plot density at approximately $d = 0.00$. We chose to estimate four intervals because the number of effect sizes per group in both data sets is relatively small, and with more than four intervals, some intervals contained fewer than 10 effect sizes, making estimation of the weights difficult. One typically wants at least 10 to 15 effects observed within each interval to adequately estimate the weights (Vevea & Woods, 2005). Using these intervals ensured that all intervals contained more than 10 effect sizes, with some containing as many as 30. This produced three adjusted models, one for the full data set and one for each group, with three adjusted mean and variance component estimates.

We used likelihood ratio tests to compare these models. It is possible to compare each adjusted model to its corresponding unadjusted mean model with a likelihood ratio test because the unadjusted mean-only model is equivalent to a special case of the adjusted model with all weights fixed to one, where all p values are equally likely to survive the selection process (Vevea & Hedges, 1995). In this way, the likelihood ratio tests comparing the adjusted models to the unadjusted models indicate whether allowing the weights to vary across intervals represents the data more accurately than fixing them all to one (the case of no publication bias). A significant likelihood ratio test indicates that there is a need for a selection model (publication bias may be present). The test statistic is calculated as twice the difference in the log-likelihoods of the models being compared, and has df equal to the number of parameters being constrained; in this case, each adjusted model estimates five parameters (a mean, a variance component, and three weights) and each unadjusted model estimates two (a mean and variance component), so these likelihood ratio tests have $df = 3$.

We then created what we call an "adjusted combined model" for each data set by summing the likelihoods of the adjusted models for both levels of the respective study characteristic. This adjusted combined model is equivalent to estimating the weight-function model with both a mean model (allowing the mean effect to vary across groups by estimating two means) and two sets of weights for p value intervals—that is, a model that allows both the mean *and* the weights to vary across groups. We tested this model against both the adjusted model for the entire data set (the "adjusted full model") and the adjusted full model with a moderator incorporated. The first likelihood ratio test is equivalent to testing a model that allows both the mean and the weights to vary against a model that allows neither to vary and estimates only one set of weights. In cases where the population mean of each group differs regardless of publication bias, this test can become significant due to the addition of a mean model even if publication bias is not actually a threat. Therefore, we implemented the second likelihood ratio test to compensate for this confound. This second test, comparing the adjusted combined model against the adjusted full model with a moderator, is equivalent to testing a model that allows the mean *and* the weights to vary against a model that allows only the mean to vary, estimating one set of weights for the entire data set. If this likelihood ratio test is significant, it indicates that allowing the weights to vary adds useful information to the model and represents the data more accurately than estimating only one set of weights, which in turn indicates that knowledge is gained by allowing patterns of publication bias to vary across groups.

Although the R code to implement the Vevea and Hedges (1995) weight-function model is not publicly available, the authors have access and are working on releasing the model to the public in the form of a *Shiny* web application (RStudio, 2012).[5] For those interested in experimenting with other weight-function models for publication bias, there is an R package called *selectMeta* (Rufibach, 2011) available via CRAN that can implement the Iyengar and Greenhouse (1988) and Dear and Begg (1992) weight-function models.

Results

Hyde and Linn (1988): "Gender Differences in Verbal Ability: A Meta-Analysis"

The Hyde and Linn (1988) meta-analysis attempted to determine whether gender differences in verbal ability favoring females were a well-established psychological finding. They found that

[4] Other weight-function models for publication bias include those proposed by Copas and Shi (2000), Dear and Begg (1992); Hedges (1992), Iyengar and Greenhouse (1988), and Lane and Dunlap (1978).

[5] *Shiny* is a web application framework for R, created by RStudio (2012). It is free and requires no advanced web development skills. For more information, visit http://shiny.rstudio.com. Our application is currently online at https://vevealab.shinyapps.io/WeightFunctionModel (case-sensitive), although still in beta. We invite any users who encounter bugs or errors while using the application to contact us at kcoburn@ucmerced.edu or jvevea@ucmerced.edu.

these gender differences decreased post-1974 and questioned whether the change might be related to a book published by Maccoby and Jacklin (1974) which pointed out a bias against nonsignificant research.

We obtained the effect sizes from the Hyde and Linn (1988) meta-analysis and divided them into two groups based on year of publication: those published in 1973 or earlier ($k = 52$) and those published in 1974 or later ($k = 69$). A table in the article provided the effect sizes, along with the total sample size for each effect. Because the sample size per group was not provided, we assumed that half the total N were males and the other half females, and calculated the sampling variances, v, based on that assumption.[6] The authors mention excluding two effect sizes from their analyses; they specify that the first effect, from Ramist and Arbeiter (1986, as cited in Hyde & Linn, 1988), came from data on 977,361 participants who took the SAT in 1985, and they excluded it because its precision was so extreme that it affected all their analyses. We excluded the Ramist and Arbeiter (1986) effect as well, leaving us with 120 effect sizes. We did not exclude the other effect size because it was not from such a large study, it was not easily identifiable from the information in the article, and we did not wish to keep our analyses completely identical to the original.

We estimated three mean-only random-effects models using the *rma* function provided in metafor (Viechtbauer, 2010) to obtain mean estimates for the full data set, the studies published in 1973 or earlier, and the studies published in 1974 or later. Table 1 provides estimates of these means and variance components; β_0 represents the mean and τ^2 represents the variance component. τ^2 is a measure of the amount of between-studies heterogeneity present in the data. We obtained a mean of $d = 0.14$ for the full data set ($\tau^2 = 0.06$), a mean of $d = 0.16$ ($\tau^2 = 0.09$) for the earlier studies, and a mean of $d = 0.12$ ($\tau^2 = 0.03$) for the later studies. These mean differences vary somewhat from those obtained as weighted averages in the original article. Again, however, this pattern indicates that studies more recently published found a smaller average effect. We went on to use the five methods described above to assess whether these methods indicated that the pattern of publication bias, or the selection model, differed across these two groups of studies.[7]

Funnel plots. Figure 2 presents a contour-enhanced funnel plot of all effect sizes ($k = 120$) against their standard errors. With the exception of one effect size near $d = 1.50$, most effect sizes fall in the range of $d = -1.00$ to $d = 1.00$, and the majority fall between $d = 0.00$ and $d = 0.50$. At the point where effect sizes become negative ($d = 0.00$), there is a drastic drop in density, such that very few effect sizes are negative and the vast majority are positive, indicating a general female superiority in terms of verbal ability. Such a sharp change in density is unlikely to occur naturally in an unbiased distribution of estimated effects. There are also many positive effect sizes that fall within the bounds of the significant contour lines, but very few that are both negative and significant. This also indicates extreme prejudice against negative effects.

Table 1
Hyde and Linn (1988) Random-Effects Meta-Analyses

Data set	β_0	τ^2	p
Full	0.141	0.059	.000
Early	0.158	0.086	.000
Later	0.116	0.026	.000

Note. Maximum-likelihood estimates are presented. β_0 represents the mean; τ^2 represents the variance component.

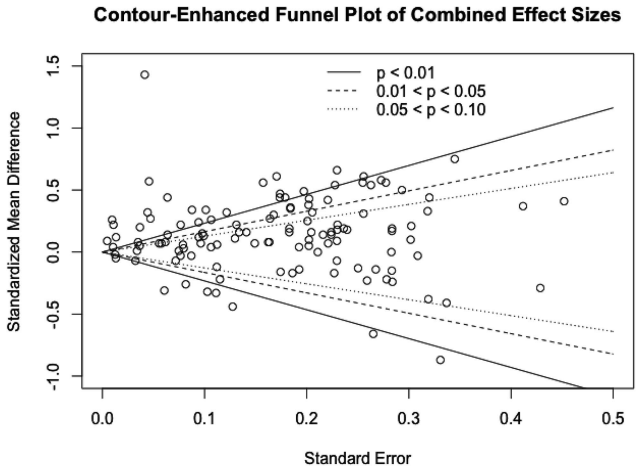

Figure 2. Hyde and Linn (1988) contour-enhanced funnel plot of combined effect sizes.

Figure 3 and Figure 4 provide the contour-enhanced funnel plots for pre- and post-1974 studies. Both display some signs of asymmetry, and in both cases the asymmetry takes the form of a drop in density where effect sizes become negative ($d = 0.00$). The group of effect sizes published in 1973 or earlier (here denoted the "earlier" group), however, has a more drastic reduction in density than those published in 1974 or later (denoted the "later" group). The funnel plot of the earlier effect sizes shows only a very few effects below the $d = 0.00$ mark, whereas the funnel plot of the later effect sizes shows a much greater percentage of effects (almost a third, based on visual approximation) below the $d = 0.00$ mark. In addition, the pattern of fewer negative effects falling in the significant contour lines continues for the earlier effects, but weakens for the later effects.

A visual assessment of funnel plots alone appears to confirm the idea that a different selection function exists for the two cases. The pattern of asymmetry changes if one shifts one's attention from the funnel plot for the overall data set to the individual funnel plots for each level of the study characteristic; the earlier effect sizes are more positive overall than the later ones.

Trim and fill. Table 2 lists the results of the trim-and-fill analyses. For the full data set, the L_o estimator added 10 effect sizes on the left side of the plot (Figure 5 presents a funnel plot showing these added effects). After imputing 10 effects, the estimate is adjusted to $d = 0.10$. The R_o estimator did not add any effects to the full data set. For the early studies, the L_o estimator

[6] This assumption is doubtless untrue, but serves the purpose of illustrating our methods.

[7] To allow for publication lag, we repeated all analyses to compare those effect sizes published pre- and post-1975. These additional analyses produced extremely similar results. No significance tests or factors that affected interpretation were changed. As such, only the results comparing effect sizes pre- and post-1974 are featured here.

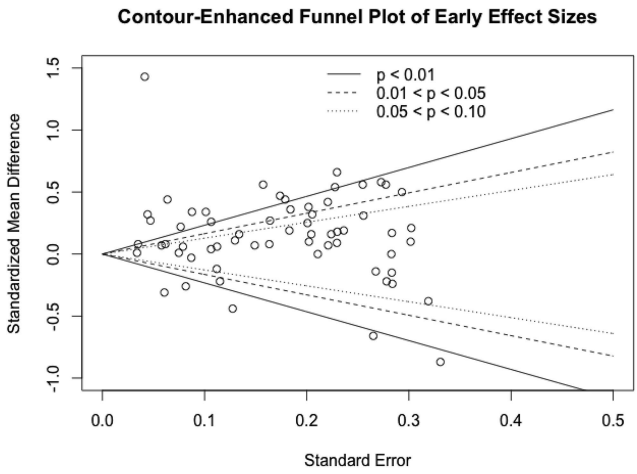

Figure 3. Hyde and Linn (1988) contour-enhanced funnel plot of effect sizes published, 1973 or earlier.

added 29 effect sizes on the right side of the plot, increasing the average effect from $d = 0.224$ to $d = 0.540$, or by about one third of a standard deviation. Note that the R_0 estimator did not add any effect sizes here, although simulations have indicated it is more accurate if k is larger (Duval & Tweedie, 2000b). Note also that the added effect sizes *increased* the mean effect, even though the funnel plots seem to indicate that the adjusted effect should decrease. This occurred regardless of the direction we specified in estimation; even guiding the function to the other side of the plot did not result in an adjustment. For the later effect sizes, the L_0 estimator added four effects to the left side of the plot. This adjusted the mean estimate downward to $d = 0.100$.

These analyses may provide some support for the theory that the effect sizes published earlier, before the release of Maccoby and Jacklin (1974), were more prone to publication bias than those published after the book's release; they are, however, somewhat unclear. The L_0 estimator added effects to all three groups, and the R_0 estimator added no effect sizes at all.

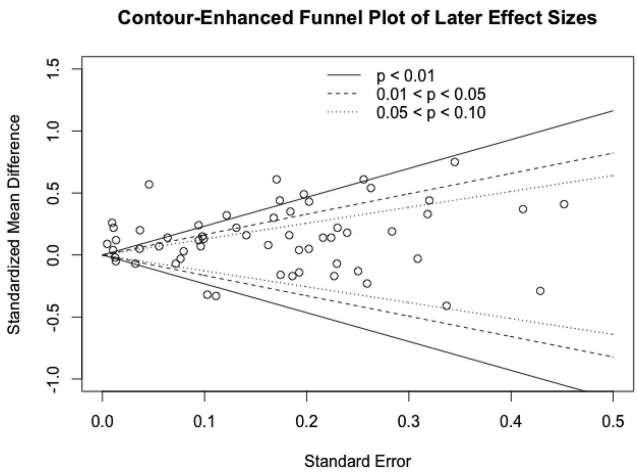

Figure 4. Hyde and Linn (1988) contour-enhanced funnel plot of effect sizes published, 1974 or later.

Table 2
Trim-and-Fill Analyses for Hyde and Linn (1988)

Data set, estimator, and side	k	k missing	β_0	New β_0	Difference in β_0
Full					
L_0					
Left	121	10	0.111	0.101	0.024
Right	121	0	0.111	0.111	0.000
R_0					
Left	121	0	0.111	0.111	0.000
Right	121	0	0.111	0.111	0.000
Early					
L_0					
Left	62	0	0.224	0.224	0.000
Right	62	29	0.224	0.540	0.316
R_0					
Left	62	0	0.224	0.224	0.000
Right	62	0	0.224	0.224	0.000
Later					
L_0					
Left	58	4	0.103	0.100	0.003
Right	58	0	0.103	0.103	0.000
R_0					
Left	58	0	0.103	0.103	0.001
Right	58	0	0.103	0.103	0.000

Note. All effect sizes were estimated with a fixed-effects model. β_0 represents the estimated mean effect. New β_0 is the new estimate after adjustment by trim and fill. The column "Difference in β_0" represents the absolute value of the difference.

Table 3 presents the results of the trim-and-fill analysis with a metaregression algorithm. Here, the earlier studies were coded 0 and the later studies 1. The L_0 estimator imputes effects, but the R_0 estimator imputes no effects. Using MLE, the L_0 estimator imputes approximately the same number of effects at each level (4 to the earlier studies and 6 to the later ones). Using MOM, however, it imputes 10 effects to the earlier studies and none to the later studies. This indicates that differing patterns of publication bias may be a concern, such that the earlier studies appear to be subject to bias and the later studies do not. Recall the unadjusted mean estimates for the earlier studies ($d = 0.17$) and the later studies

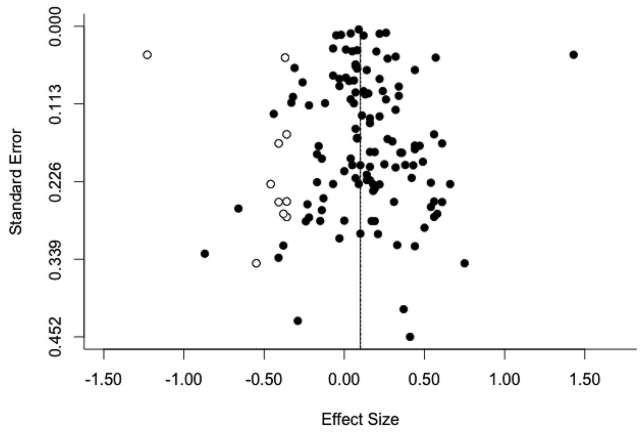

Figure 5. Hyde and Linn (1988) trim-and-fill plot of full data set, L estimator, left side. Observed effects are represented by filled-in circles; added effects are represented by empty circles.

($d = 0.11$) from Table 1. The adjusted estimates obtained with the L_0 estimator, using MLE, are $d = 0.13$ for the earlier studies and $d = 0.06$ for the later studies. Using MOM, they are also $d = 0.13$ for the earlier studies and $d = 0.06$ for the later studies. Both of these estimates have been adjusted downward, as is consistent with a visual inspection of the funnel plots.

Egger's linear regression. Table 4 provides the results of the Egger's linear regression analyses. The test on the full data set was significant ($Z = 3.43$, $p < .05$). The test on the earlier studies was also significant ($Z = -4.56$, $p < .05$). The test on the later studies was nonsignificant ($Z = 0.85$, $p = .40$). These results provide additional evidence that publication bias affects the earlier studies and may not affect the later ones.

Cumulative meta-analysis. The cumulative meta-analyses of the complete data set, the earlier studies, and the later studies are presented in Figures 6 and 7, respectively.[8] Drift to the right, for effects sorted by precision, indicates that smaller studies with smaller effects are being suppressed by publication bias, resulting in a bias favoring large, significant effects (McDaniel, 2009). As those small studies with large effects are added to the meta-analysis, the cumulative meta-analytic estimate of the mean effect drifts to the right. In Figure 6, the cumulative mean shifts from $d = 0.01$ to $d = 0.22$. There is a leap in precision from the second effect size to the third, which could represent some drift to the right but is somewhat difficult to interpret as there are virtually no signs of drift later in the plot. For the later studies, in Figure 7, the cumulative mean shifts from $d = 0.09$ to $d = 0.10$. There is no sign of drift at all.

Weight-function model. Table 5 presents the results of the weight-function model analyses for the full data set and both individual groups. The mean and variance components, adjusted for publication bias, are presented under β_0 and τ^2, respectively, and can be compared with the unadjusted estimates in Table 1. The weight for the first interval, $p < .05$, is fixed to 1, and the weights for the other intervals should be interpreted relative to the first interval. For example, Weight 4 for the earlier studies is estimated at 0.188, indicating that effect sizes published before 1974 with a p value between .50 and 1.00 are less than 0.20 times as likely to survive the selection process as significant effect sizes. The likelihood ratio tests in the subsequent columns compare the unadjusted model for each group (presented in Table 1) to the adjusted model presented here. A significant test indicates that the adjusted model fits the data significantly better.

Table 3
Trim and Fill With Metaregression Analyses for Hyde and Linn (1988)

Estimator and method	β_0	β_1	τ^2	k early studies	k later studies
L_0					
MLE	0.130	−0.071	0.040	4	6
MOM	0.127	−0.065	0.040	10	0
R_0					
MLE	0.164	−0.048	0.025	0	0
MOM	0.164	−0.048	0.025	0	0

Note. Estimates provided are adjusted random effects. k are the number of effect sizes imputed to each level of the study characteristic. β_0 represents the mean of the group coded 0; here, the earlier studies. The sum of β_0 and β_1 represents the mean of the group coded 1; here, the later studies. τ^2 represents the variance component. MLE = maximum likelihood; MOM = method of moments.

Table 4
Egger's Regression (Hyde & Linn, 1988)

Data set	Z	p
Full	3.434	.001
Early	−4.562	.000
Later	0.854	.393

Note. Z statistics test for funnel plot asymmetry. All tests use fixed-effects models predicted by standard error.

The likelihood ratio test for the full data set was nonsignificant, indicating that estimating the adjusted model did not add information, $\chi^2(3) = 7.17$, $p = .13$. The overall mean estimate was adjusted downward, from $d = 0.14$ to $d = 0.04$. The likelihood ratio test for the effect sizes published earlier than 1974, the "earlier" effects, was significant, indicating the presence of publication bias, $\chi^2(3) = 12.15$, $p < .05$. The weight-function model also adjusted the mean estimate for the earlier effect sizes downward from $d = 0.17$ to $d = -0.07$, which is consistent with our previous observations of the funnel plot. The likelihood ratio test for the effect sizes published after 1974, the later effects, was nonsignificant, indicating that the model unadjusted for publication bias does an equally good job of representing the data, $\chi^2(3) = 1.45$, $p = .84$. Yet again, however, the adjusted model reduced the mean estimate for the "later" studies, from $d = 0.11$ to $d = 0.03$.

Next, a likelihood ratio test compared the adjusted full model and the adjusted combined model. Recall that we calculated the likelihoods of the combined model by summing the likelihoods of the two individual models. For this test, $df = 5$, the number of parameters constrained (10 parameters were estimated for the adjusted combined model, and five were estimated for the combined unadjusted model). Table 6 presents the results of this test. This likelihood ratio test was significant, $\chi^2(5) = 16.89$, $p < .01$, indicating that the adjusted combined model (estimating a mean model and a different set of weights for each group) represents the data more accurately than estimating one weight-function model for the complete data set without a mean model. A likelihood ratio test comparing the adjusted full mean model with the adjusted combined model was also significant, $\chi^2(4) = 16.49$, $p < .01$, indicating that some information is gained by representing two different patterns of publication bias.

Summary. The Hyde and Linn (1988) analyses indicate that publication bias affects the earlier studies, published before 1974, but does not appear to affect the later studies, published after 1974, to the same degree.

Funnel plots are a purely visual assessment and as such are inherently ambiguous; however, the funnel plot of the earlier

[8] Although we sorted these cumulative meta-analyses by precision, as recommended in the literature, in this case it would also be possible to sort the studies by year of publication. We assessed a plot sorted by year, but in this case it is difficult to interpret. There are several large jumps in the plot due to sharp increases in precision from large samples. As such, we did not include it in the article, but would like to note that this could prove to be a useful assessment in similar cases.

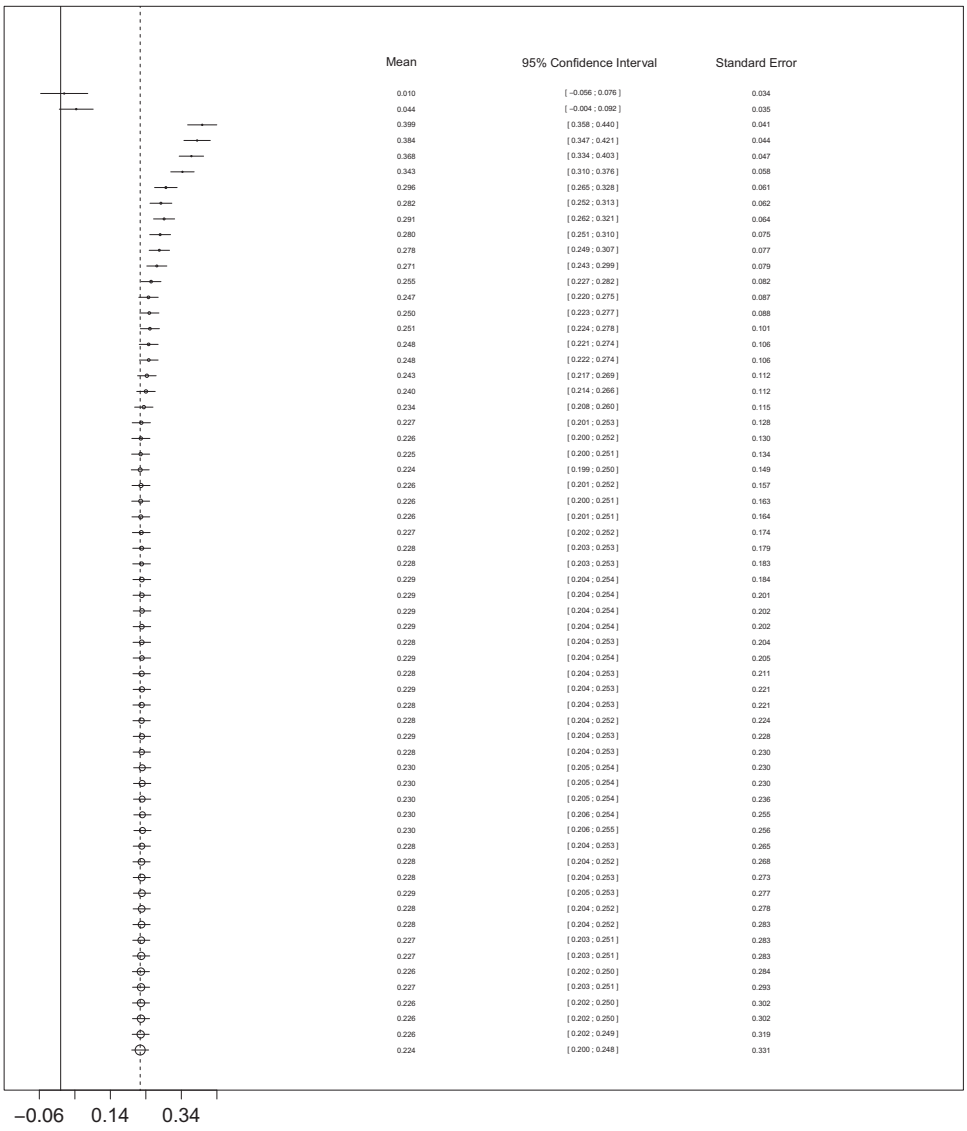

Figure 6. Hyde and Linn (1988) cumulative meta-analysis by precision, early effect sizes. "Mean" is the cumulative meta-analytic mean. "95% Confidence Interval" is the cumulative 95% confidence interval. "Precision" is the precision of each individual study added, or 1/standard error. Larger circles represent effect sizes with larger standard errors. The solid vertical line represents $d = 0.00$; the dashed line is drawn at the final cumulative mean.

studies does show more signs of asymmetry than that of the later studies. Both variants of trim and fill, with metaregression and without, also indicate that the earlier studies are subject to bias. For trim and fill without metaregression, one estimator imputes 29 effect sizes to the earlier studies, whereas only four effects are imputed to the later studies. For trim and fill with metaregression, only the L estimator imputes effect sizes; it imputes 4 to the early group and 6 to the later group using maximum likelihood estimation, and it imputes 10 to the early group using method of moments estimation. Egger's regression is significant for the complete data set and for the earlier studies ($p < .05$), but nonsignificant for the later studies. Among the cumulative meta-analyses sorted by precision, none of the plots shows clear drift to the right. Last, likelihood ratio tests using the weight-function model indicates that significant information is gained by estimating two different sets of weights, one for the earlier studies and one for the later studies; the early studies are better represented by the adjusted model ($p < .05$), and the later studies are not.

Overall, the later studies do not seem subject to publication bias, but the earlier studies do. This indicates that a change in expectations about gender differences that coincided with the release of the pivotal book by Maccoby and Jacklin (1974) actually may have influenced the publication of nonsignificant results, and that the mean difference between the two groups may be an artifact of publication bias.

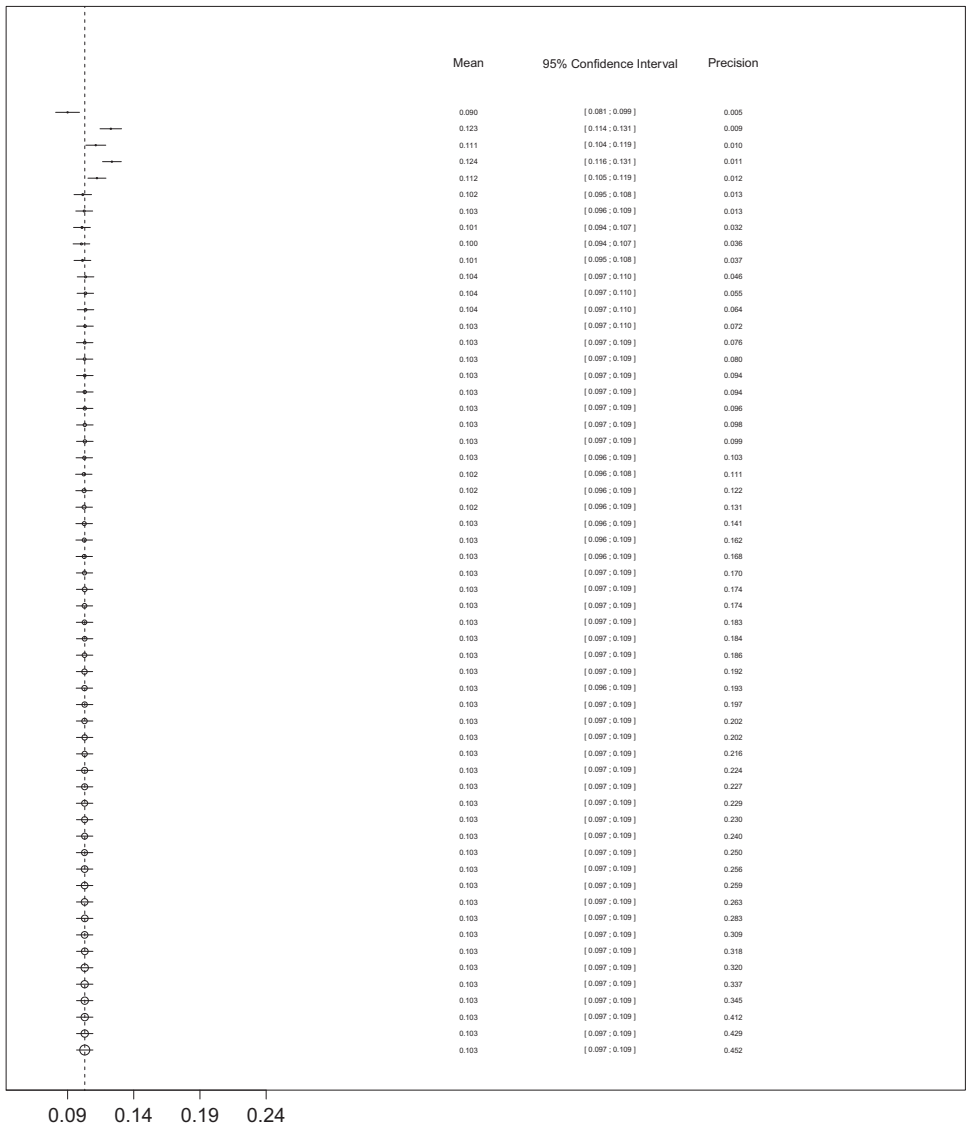

Figure 7. Hyde and Linn (1988) cumulative meta-analysis by precision, later effect sizes. "Mean" is the cumulative meta-analytic mean. "95% Confidence Interval" is the cumulative 95% confidence interval. "Precision" is the precision of each individual study added, or 1/standard error. Larger circles represent effect sizes with larger standard errors. The solid vertical line represents $d = 0.00$; the dashed line is drawn at the final cumulative mean.

Kling et al. (1999): "Gender Differences in Self-Esteem: A Meta-Analysis"

This meta-analysis assessed gender differences in self-esteem and found a significant mean difference between effect sizes that were the primary focus of a study and effect sizes that were the secondary focus of a study (referred to here as *gender-focused* or *non-gender-focused*). The authors theorized that this might be because effect sizes from gender-focused studies were more likely to be subject to publication bias than those from non-gender-focused studies.

A table in the article provided the data set, but without the moderator representing centrality of gender. Consequently, we coded the articles again following the protocol the authors provided. Articles met the definition of *gender-focused* if either (a) the abstract described gender comparisons or (b) the title or abstract included words such as *sex*, *gender*, *masculine*, or *feminine* (Kling et al., 1999). After recoding, 109 effect sizes were identified as gender-focused, 87 were identified as non-gender-focused, and the remaining 20 effect sizes were omitted from the analysis because they did not mention gender in their titles and were from sources without abstracts. This resulted in a total of 196 effect sizes. There were some differences in coding in that we coded 10 more effect sizes as gender-focused and 10 fewer as non-gender-focused, but because the authors did not provide their codes, we could not assess these differences.

Table 5
Weight-Function Model Analyses for Hyde and Linn (1988)

Data set	β_0	τ^2	Weight 1	Weight 2	Weight 3	Weight 4	LR χ^2	*df*	*p*
Full	0.035	0.070	1	0.912	0.926	0.380	7.167	3	.127
Early	−0.065	0.116	1	0.650	1.030	0.188	12.152	3	.016
Later	0.027	0.084	1	1.075	0.637	0.674	1.445	3	.836

Note. Parameter estimates provided are for the adjusted model. The first step is fixed to one for estimation purposes. The likelihood ratio test is performed comparing the adjusted and unadjusted models. β_0 represents the mean; τ^2 represents the variance component. Weight 1 matches the interval $p < .05$; Weight 2, the interval $0.05 < p < .20$; Weight 3, the interval $0.20 < p < .50$; and Weight 4, the interval $0.50 < p < 1.00$. All weights are interpreted relative to the interval $p < .05$, which is fixed to 1.

We estimated three random-effects models to get unadjusted estimates of the mean effect for the full data set, the gender-focused studies, and the non-gender-focused studies. Table 7 provides estimates of these means and variance components; β_0 represents the mean and τ^2 represents the variance component. τ^2 is a measure of the amount of between-studies heterogeneity present in the data. We obtained an average effect of $d = 0.21$ ($\tau^2 = 0.02$) for the full data set, an effect of $d = 0.23$ ($\tau^2 = 0.02$) for the gender-focused studies, and an effect of $d = 0.11$ ($\tau^2 = 0.02$) for the non-gender-focused studies. The pattern indicates that effect sizes from non-gender-focused studies were smaller on average. We then assessed publication bias using funnel plots, trim and fill (with and without a metaregression algorithm), Egger's linear regression, cumulative meta-analysis, and the weight-function model.

Funnel plots. Figure 8 presents the funnel plot of all effect sizes ($k = 196$) against their standard errors. Most effect sizes fall between approximately $d = 0.00$ and $d = 0.60$, indicating an overall positive effect size. The effect sizes are clustered in the upper left corner of the plot, between approximately $d = 0.00$ and $d = 0.40$, with standard errors ranging from approximately 0.05 to 0.20. Few effect sizes are negative, and there appears to be a sharp drop in plot density after $d = 0.00$, creating the impression of asymmetry in that corner of the plot. An examination of the contour lines shows that more positive effect sizes than negative effect sizes are significant.

Figure 9 presents the funnel plot of the effect sizes from gender-focused studies. Again, most of the effect sizes fall between approximately $d = 0.00$ and $d = 0.50$, and they are clustered in the upper left corner, with a sharp drop in density at the $d = 0.00$ mark—a drop that is even sharper than in the overall funnel plot. Only about 10 effect sizes appear to be negative. In contrast, the funnel plot of the non-gender-focused effect sizes (Figure 10) still shows a pattern of greater density in the upper left corner, but the division at $d = 0.00$ is less sharp and more effect sizes are negative, with some ranging as low as $d = -0.40$. The non-gender-focused effect sizes appear more spread out than the gender-focused ones. The contour lines show that the gender-focused effects continue the pattern of more positive significant effects than negative ones; the non-gender-focused effects are slightly more evenly distributed. However, in comparison to the funnel plots from the Hyde and Linn (1988) meta-analysis, differing patterns of asymmetry are not as clear. A visual assessment of the funnel plots appears to confirm the idea that the pattern of publication bias differs here across levels of centrality of gender.

Trim and fill. Table 8 presents the results of the trim-and-fill analyses. Both estimators added effect sizes to the complete data set, although the L_0 estimator added only four (not pictured) and the R_0 estimator added 10 (pictured in Figure 11). The R_0 estimator adjusted the mean estimate for the complete data set to $d = 0.22$ (from $d = 0.21$). For individual assessments of the groups, neither estimator added more than five effect sizes; the R_0 estimator added two effect sizes in both groups, and the L_0 estimator added five effect sizes for the gender-focused studies and two for the non-gender-focused studies. The adjusted mean estimate for both groups did not change by more than 0.07 (with the L_0 estimator). Most added effect sizes were in the upper halves of the funnel plots, ranging from about $d = 0.40$ to $d = 0.80$, and increased the mean effect size for the corresponding group.

This pattern of results indicates that publication bias may be more severe in the combined data set than in the individual groups of effect sizes. Trim and fill added one more effect size to the non-gender-focused studies than the gender-focused studies, but this difference is negligible. Based on the trim-and-fill results alone, the pattern of publication bias does not appear to differ across groups. However, the differing results of these estimators should call this finding into question. In addition, note that both estimators, if they did add effect sizes, tended to add these data points in the upper halves of the funnel plots, although visual

Table 6
Likelihood-Ratio Tests for Hyde and Linn (1988)

Model 1	Model 2	LR χ^2	*df*	*p*
Adjusted full model	Adjusted combined model	16.890	5	.005
Adjusted full with mean model	Adjusted combined model	16.488	4	.002

Note. The likelihood values for the combined models were formed by summing the likelihoods of both small-group models. The degrees of freedom are the difference in the number of parameters estimated.

Table 7
Kling et al. (1999) Random-Effects Meta-Analyses

Data set	β_0	τ^2	*p*
Full	0.205	0.018	.000
GF	0.232	0.015	.000
NGF	0.159	0.020	.000

Note. Maximum-likelihood estimates are presented. β_0 represents the mean; τ^2 represents the variance component. GF = gender-focused effects; NGF = non-gender-focused effects.

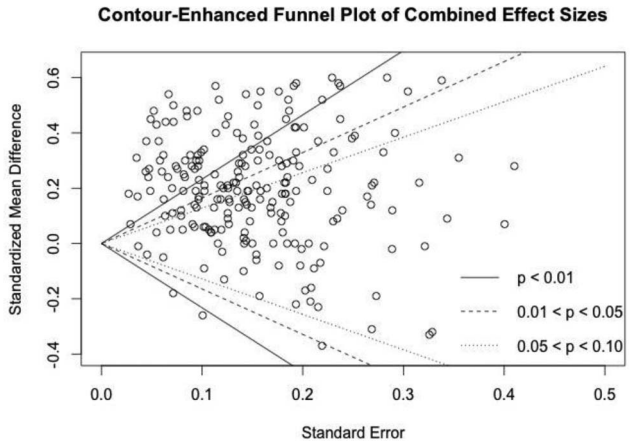

Figure 8. Kling, Hyde, Showers, and Buswell (1999) contour-enhanced funnel plot of combined effect sizes.

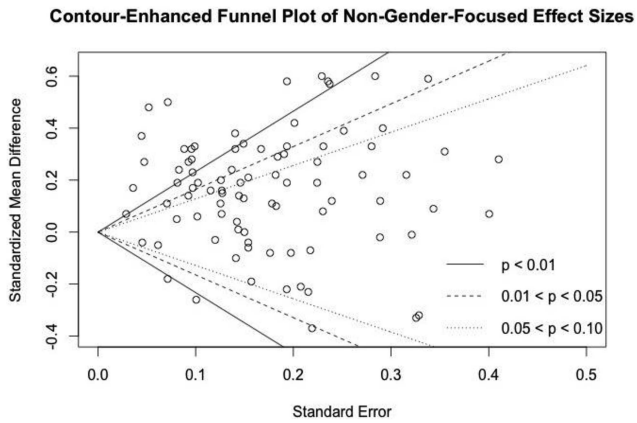

Figure 10. Kling, Hyde, Showers, and Buswell (1999) funnel plot of non-gender-focused effect sizes.

assessments indicate a lack of effect sizes in the lower halves. Again, this occurred regardless of which side of the plot was specified in the analysis.

Table 9 presents the results of the trim-and-fill analysis with a metaregression algorithm. Here, the gender-focused studies were coded 0 and the non-gender-focused studies coded 1. Neither the L_o estimator nor the R_o estimator imputed any effects at all. This does not provide evidence that publication bias is a concern.

Egger's linear regression. Table 10 provides the results of the Egger's linear regression analyses. Again, all three regression tests were nonsignificant, for the full data set and for both the gender-focused and non-gender-focused groups of effect sizes (p = .29, 0.60, and 0.94, respectively). These results do not provide evidence that publication bias is a concern, and do not indicate that the pattern of publication bias differs across groups.

Cumulative meta-analysis. The cumulative meta-analyses of the gender-focused studies and the non-gender-focused studies are presented in Figures 12 and 13, respectively. Neither of these figures shows any clear sign of drift to the right; neither provides evidence that publication bias is a concern. For the gender-focused effect sizes, the cumulative mean moves from d = 0.31 to d = 0.24. For the non-gender-focused effect sizes, the cumulative mean moves from d = 0.07 to d = 0.16.

Weight-function model. Table 11 presents the results of the weight-function model analyses for the full data set and both the gender-focused and non-gender-focused groups. The mean and variance components, adjusted for publication bias, are presented under β_0 and τ^2, respectively, and can be compared with the unadjusted estimates in Table 7. The weight for the first interval, $p < .05$, is fixed to 1, and the weights for the other intervals should be interpreted relative to the first interval. For example, Weight 4

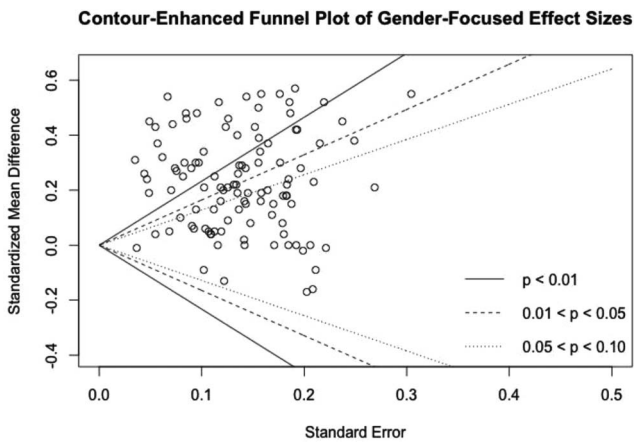

Figure 9. Kling, Hyde, Showers, and Buswell (1999) contour-enhanced funnel plot of gender-focused effect sizes.

Table 8
Trim-and-Fill Analyses for Kling et al. (1999)

Data set, estimator, and side	k	k missing	β_0	New β_0	Difference in β_0
Full					
L_0					
Left	216	0	0.207	0.207	0.000
Right	216	4	0.207	0.208	0.001
R_0					
Left	216	0	0.207	0.207	0.000
Right	216	10	0.207	0.212	0.005
GF					
L_0					
Left	109	0	0.237	0.237	0.000
Right	109	5	0.237	0.244	0.007
R_0					
Left	109	0	0.237	0.237	0.000
Right	109	2	0.237	0.239	0.002
NGF					
L_0					
Left	87	2	0.161	0.159	0.002
Right	87	0	0.161	0.161	0.000
R_0					
Left	87	0	0.161	0.161	0.000
Right	87	2	0.161	0.163	0.002

Note. All effect sizes were estimated with a fixed-effects model. β_0 represents the estimated mean effect. New β_0 is the new estimate after adjustment by trim and fill. The column of "Difference in β_0" represents the absolute value of the difference. GF = gender-focused effects; NGF = non-gender-focused effects.

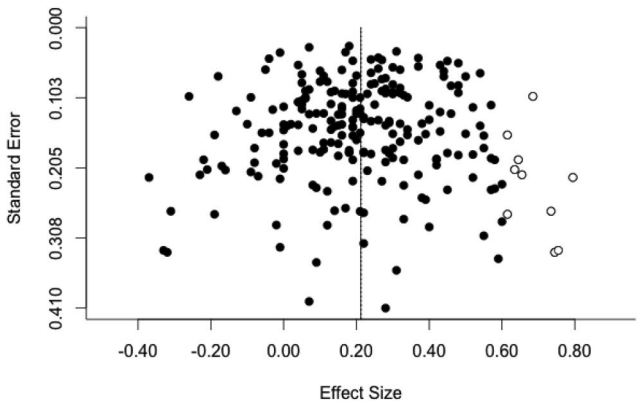

Figure 11. Kling, Hyde, Showers, and Buswell (1999) trim-and-fill plot of full data set, R estimator, right side. Filled-in circles represent observed effects; empty circles represent imputed effects.

for the gender-focused studies is estimated at 2.456, indicating that effect sizes from gender-focused studies with a p value between .50 and 1.00 are more than twice as likely to survive the selection process as significant effect sizes. The likelihood ratio tests in the subsequent columns compare the unadjusted model for each group (presented in Table 7) to the adjusted model presented here. A significant test indicates that the adjusted model fits the data significantly better.

The likelihood ratio test comparing the unadjusted model for the complete data set to the adjusted model was nonsignificant, $\chi^2(3) = 2.725$, $p = .44$. The overall mean estimate was adjusted upward, from $d = 0.21$ to $d = 0.24$. Based on the results of this comparison alone, it appears that publication bias is not a concern. For the gender-focused effects, the likelihood ratio test was also nonsignificant, $\chi^2(3) = 2.731$, $p = .44$) and the mean was adjusted upward, from $d = 0.23$ to $d = 0.27$. For the non-gender-focused effects, the likelihood ratio test also did not detect a significant difference between the adjusted and unadjusted models, $\chi^2(3) = 1.746$, $p = .63$, and the mean effect was adjusted upward, but still rounds to $d = 0.16$. These results alone do not indicate that the pattern of publication bias differs significantly across groups.

Table 9
Trim and Fill With Metaregression Analyses for Kling et al. (1999)

Estimator and method	β_0	β_1	τ^2	k GF	k NGF
L_0					
MLE	0.231	−0.072	0.020	0	0
MOM	0.231	−0.072	0.020	0	0
R_0					
MLE	0.231	−0.072	0.020	0	0
MOM	0.231	−0.072	0.020	0	0

Note. Estimates provided are adjusted random-effects.. k are the number of studies imputed to each level of the study characteristic. β_0 represents the mean of the group coded 0, here the gender-focused (GF) studies; the sum of β_0 and β_1 represents the mean of the group coded 1, here the non-gender-focused (NGF) studies. MLE = maximum likelihood; MOM = method of moments.

Table 10
Egger's Regression (Kling et al., 1999)

Data set	Z	p
Full	−1.068	.286
Gender-focused	−0.528	.598
Non-gender-focused	−0.070	.944

Note. Z statistics test for funnel plot asymmetry. All tests use fixed-effects models predicted by standard error.

Next, a likelihood ratio test compared the adjusted full model and the adjusted combined model (which we calculated, as before, by summing the likelihoods for the individual models). Again, there were five *df* for this test, equal to the number of constrained parameters. Table 12 displays these results. This test was significant, $\chi^2(5) = 41.017$, $p < .001$. These results at first appeared unusual; the adjusted combined model emerged superior from this likelihood ratio test, indicating that two sets of weights fit the data significantly better than one. If estimated individually, however, none of the adjusted models provided significantly more information than the unadjusted models. It did not seem plausible that the adjusted model for both groups combined could model the data more effectively than the unadjusted models.

The second likelihood ratio test, which compared the adjusted full mean model with the adjusted combined model, shed some light on these results. Recall that this is equivalent to testing a model allowing the mean to vary against a model that allows both the mean and the weights to vary. The results of this likelihood ratio test are presented in Table 12. This likelihood ratio test was nonsignificant, $\chi^2(4) = 3.088$, $p = .54$, indicating that representing two different patterns of publication bias did not provide significantly more information than allowing the mean to vary across groups. In other words, the first likelihood ratio test was significant because the mean effect size differs significantly across groups, and there is no evidence indicating that this difference is an artifact of publication bias.

Summary. The Kling et al. (1999) analyses indicate that neither the effect sizes from gender-focused studies nor the effect sizes from non-gender-focused studies seem to be subject to publication bias, and by this logic, that the pattern of publication bias does not differ across levels of the study characteristic.

A visual examination of the funnel plots might give the impression of some asymmetry among the gender-focused effects, but none of the other analyses support this observation. Trim and fill without a metaregression algorithm imputes a few effect sizes to both the gender-focused and the non-gender-focused studies, but trim and fill with metaregression imputes no effect sizes at all. Egger's regression is nonsignificant for both the gender-focused and the non-gender-focused studies. Among the cumulative meta-analyses sorted by precision, neither plot demonstrates drift to the right. Last, likelihood ratio tests using the weight-function model indicate that adjusting for publication bias does not provide significantly more information for either the gender-focused or the non-gender-focused studies.

In this case, there is evidence of a true mean difference between the gender-focused and the non-gender-focused studies, and this mean difference does not appear to be an artifact of bias.

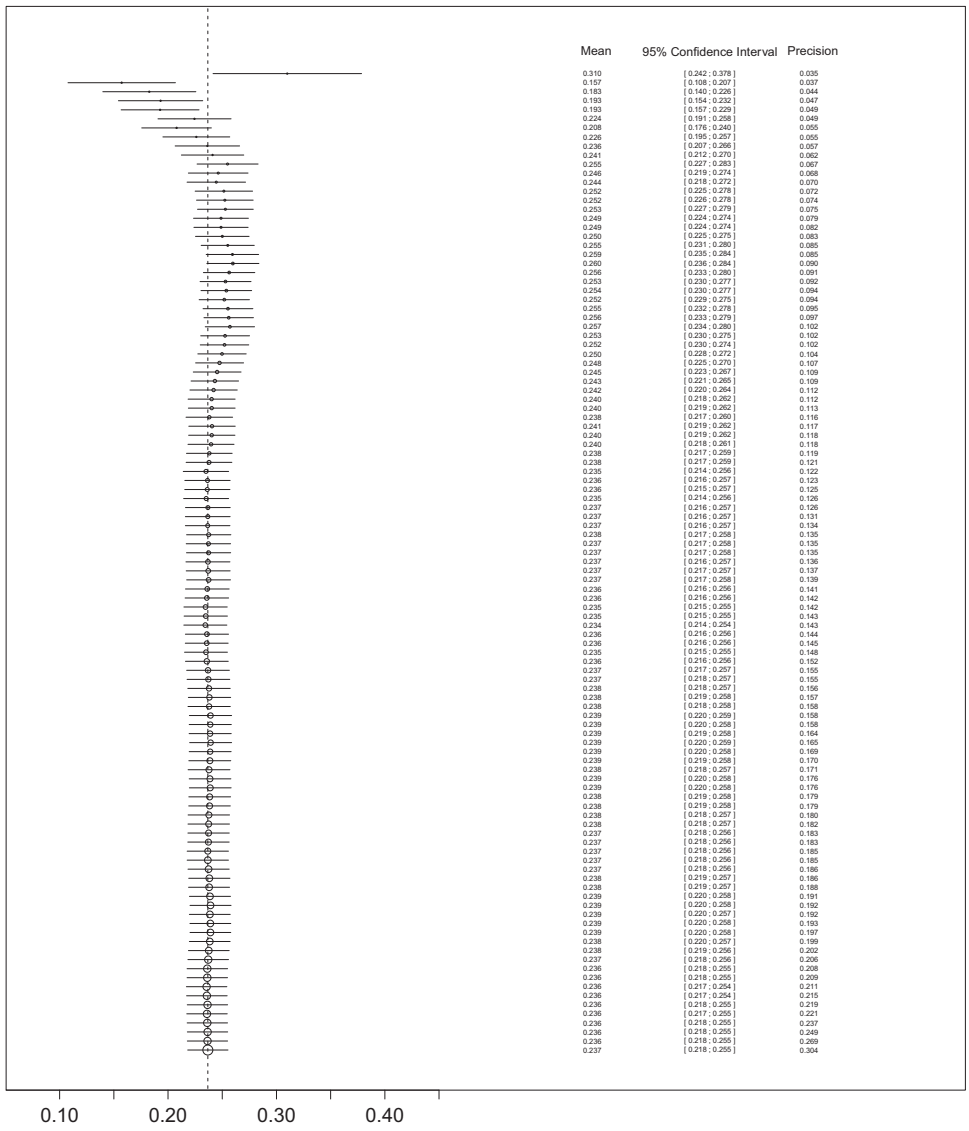

Figure 12. Kling, Hyde, Showers, and Buswell (1999) cumulative meta-analysis by precision, gender-focused effects. "Mean" is the cumulative meta-analytic mean. "95% Confidence Interval" is the cumulative 95% confidence interval. "Precision" is the precision of each individual study added, or 1/standard error. Larger circles represent effect sizes with larger standard errors. The solid vertical line represents $d = 0.00$; the dashed line is drawn at the final cumulative mean.

Discussion

In this article, we used data sets from two meta-analyses to demonstrate that, although people often think of publication bias as a function of statistical significance, it can depend on other study characteristics. We obtained data from two meta-analyses, one on gender differences in self-esteem and one on gender differences in verbal ability (Hyde & Linn, 1988; Kling et al., 1999). For each data set, we divided the effect sizes into groups according to a study characteristic. For Hyde and Linn (1988), the relevant characteristic was year of publication, before or after 1974; for Kling et al. (1999), the characteristic was the question of whether the focus of the study was primarily on gender. We then assessed publication bias three times per data set using five methods, for a total of 30 assessments, and reported the results.

First we discuss the conclusions for each data set individually. We then describe some limitations of our article and implications for future research, and end with some general recommendations for meta-analyses in the future.

Hyde and Linn (1988)

When we assessed the complete data set, only trim and fill provided some evidence that publication bias was a concern. If we divided the effect sizes into two groups ("earlier" and "later" studies) and assessed the groups individually, a different picture

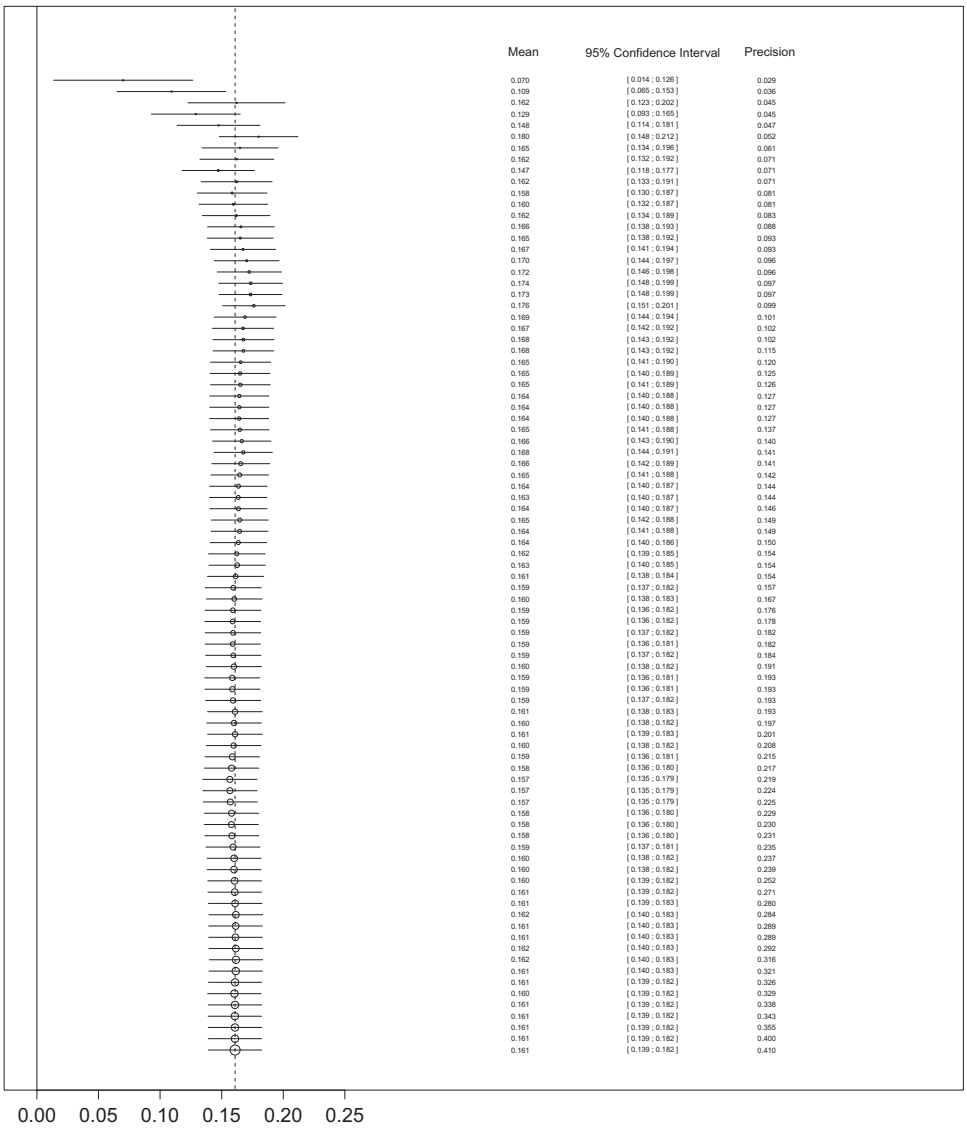

Figure 13. Kling, Hyde, Showers, and Buswell (1999) cumulative meta-analysis by precision, non-gender-focused. "Mean" is the cumulative meta-analytic mean. "95% Confidence Interval" is the cumulative 95% confidence interval. "Precision" is the precision of each individual study added, or 1/standard error. Larger circles represent effect sizes with larger standard errors. The solid vertical line represents $d = 0.00$; the dashed line is drawn at the final cumulative mean.

began to emerge. The contour-enhanced funnel plots for the earlier and later studies did appear to show a differing pattern of bias, such that the earlier studies show signs of publication bias but the later studies do not. Trim and fill indicated that publication bias might be a concern, although the results were unclear. The adapted version using a metaregression algorithm did indicate that the pattern of bias might differ; the L_0 estimator imputed effect sizes and adjusted the mean effect for both groups toward zero. Egger's regression was significant for both the full data set and the earlier studies, but nonsignificant for the later studies. Neither cumulative meta-analysis showed strong signs of drift to the right. The likelihood ratio tests comparing the weight-function models, presented in Table 6, were both significant, indicating that estimating weights for each group was significantly more informative because the pattern of bias differs significantly across the two groups. The range of mean effect sizes produced for each group varies; for the earlier studies, the mean ranges from $d = 0.01$ (with trim and fill using metaregression) to $d = 0.54$ (with the original trim-and-fill algorithm). The mean for the later studies ranges from $d = -0.05$ to $d = 0.11$ (both with trim and fill using metaregression). The unadjusted mean effect sizes were $d = 0.17$ for the earlier studies and $d = 0.11$ for the later studies.

For this meta-analysis, a study characteristic does appear to have influenced the pattern of publication bias. One plausible interpretation, as posited by Hyde and Linn (1988), is that the release of Maccoby and Jacklin's book (1974) influenced the publication of

Table 11
Weight-Function Model Analyses for Kling et al. (1999)

Dataset	β_0	τ^2	Weight 1	Weight 2	Weight 3	Weight 4	LR χ^2	*df*	*p*
Full	0.235	0.015	1	1.328	1.343	1.879	2.725	3	.436
GF	0.268	0.012	1	1.325	1.716	2.456	2.731	3	.435
NGF	0.164	0.021	1	1.339	0.889	1.143	1.746	3	.627

Note. Parameter estimates provided are for the adjusted model. The first step is fixed to one for estimation purposes. The likelihood ratio test is performed comparing the adjusted and unadjusted models. β_0 represents the mean; τ^2 represents the variance component. Weight 1 matches the interval $p < .05$; Weight 2, the interval $0.05 < p < .20$; Weight 3, the interval $0.20 < p < .50$; and Weight 4, the interval $0.50 < p < 1.00$. All weights are interpreted relative to the interval $p < .05$, which is fixed to 1. GF = gender-focused effects; NGF = non-gender-focused effects.

studies on gender differences in verbal ability. Before 1974, more studies indicated that females were superior; after 1974, this difference began to disappear.

Kling et al. (1999)

Again, when we assessed the complete data set, none of the five methods demonstrated that publication bias might be a concern. This time, though, assessing the gender-focused and non-gender-focused effects individually did not present a different picture. For both groups, neither the original trim and fill, trim and fill with a metaregression algorithm, Egger's regression, nor cumulative meta-analysis presented evidence of publication bias. Patterns of asymmetry in the funnel plots were not clear. Even the likelihood ratio tests comparing weight-function models were nonsignificant; at first, they gave the impression that the pattern of bias might differ, but ultimately the mean effect differed across groups, and allowing the weights to vary did not contribute significantly more information. The mean for the gender-focused studies ranges from $d = 0.231$, from trim and fill with metaregression, to $d = 0.268$, from the Vevea and Hedges (1995) weight-function model. The mean for the non-gender-focused studies ranges from $d = 0.152$, from trim and fill, to $d = 0.164$, from the Vevea and Hedges (1995) weight-function model. The unadjusted mean effect sizes were $d = 0.230$ for the gender-focused studies and $d = 0.110$ for the non-gender-focused studies.

This meta-analysis is an example of a case in which a suspect study characteristic does not appear to have influenced publication bias. There seems to be a significant mean difference between gender-focused and non-gender-focused studies that is not due to bias. This could be due to any number of reasons; in this case, perhaps the studies that focused primarily on gender featured more targeted interventions.

Table 12
Likelihood-Ratio Tests for Kling et al. (1999)

Model 1	Model 2	LR χ^2	*df*	*p*
Adjusted full model	Adjusted combined model	41.017	5	.000
Adjusted full with mean model	Adjusted combined model	3.088	4	.543

Note. The likelihood values for the combined models are formed by summing the likelihoods of both individual-group models.

Limitations

We encountered some difficulty selecting two meta-analyses for this article, largely because relatively few meta-analyses publish their effect sizes. In cases where meta-analysts do publish the effects, they do not always include all of them, or all of the moderators they examine. The Kling et al. (1999) data are an example of this, where our moderator of interest (focus of the study) was not included in the article. There was sufficient information provided for us to recode the studies, but we obtained slightly different results, and could not compare ours to the original. The Hyde and Linn (1988) meta-analysis also published effects but only included the total sample sizes for each study, which was not enough information to compute the sampling variances. We realize, of course, that for large meta-analyses it is not always feasible to publish all the data in the article. It would, however, be helpful for meta-analysts to release their data as supplemental information, not just for replicability but to allow other researchers to use it for demonstrations and update the analyses if more effective tools become available. Many authors have encouraged the release of data from primary studies (Freese, 2007; King, 1995; Johnson, 2001; Schneider, 2004), and researchers are beginning to call for the publication of meta-analytic effect sizes as well (American Psychological Association, 2009; Higgins, 2008; Moher et al., 1999; Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). We support this practice and feel that its implementation would benefit not only the academic community but the public.

In addition, both of the meta-analyses we examined focused on gender differences, although the study characteristics we assessed did differ. This was in part a coincidence—both of these meta-analyses were fairly large, published their data, and coded for a potentially relevant study characteristic—and in part because, as discussed in the introduction, gender differences are an area in which social preferences are shown to play a role in publication bias. We do not view this as negatively impacting generalizability, because we chose these articles solely for the purpose of demonstration; it would, however, be informative to examine more meta-analyses in different areas.

The methods we used have some limitations. To our knowledge, there have been no simulation studies examining these models' performance in a situation where the pattern of publication bias differs across groups, and therefore we know little about how well they are expected to perform in such a case. Simulations of these models' performance in this situation could be informative. In addition, the Vevea and Hedges (1995) weight-function model

requires a relatively large sample size to perform well (Vevea & Woods, 2005). Funnel plots and cumulative meta-analysis may not require such a large sample, but both are visual assessments and, as such, have an inherent degree of subjectivity.

Implications

Overall, these results suggest that publication bias that differs according to study characteristics can be a problem. The results highlight the importance of using multiple methods to assess publication bias, and of doing so across the levels of suspect study characteristics. This procedure allows meta-analysts to investigate suspect mean differences and assess whether these differences are an artifact of publication bias. Finally, they demonstrate that there is a need for a model of publication bias capable of accounting for study characteristics.

Using multiple methods to assess publication bias is related to the concept of triangulation, or the use of "multiple reference points to locate an object's exact position" (Jick, 1979, p. 602). For meta-analysis, and particularly for publication bias, triangulation consists of using multiple methods of assessment and reporting the range of results, rather than relying on one method and one point estimate (Kepes, Banks, McDaniel, & Whetzel, 2012). It is not widely implemented, as studies show only 25% of meta-analysts use two publication bias assessments, and only 3% use more than two. Even these percentages are likely inflated because 30% use some variation of the fail-safe-*N* technique, which we intentionally excluded because it is subject to a number of statistical flaws and does not provide a valid assessment of bias (Becker, 2005; McDaniel et al., 2006). Looking at the results of all five methods for only the complete data sets, a meta-analyst might not conclude that publication bias was a concern at all. This is worrisome, especially in light of the fact that many meta-analysts only use *one* of these methods, and some use none at all. If we were to assess only the complete data sets with only one of these methods, disregarding the role of study characteristics entirely, most of these methods would not indicate that publication bias was a concern. At this point, meta-analysts might stop assessing bias and assume that they could trust their results.

We also believe it is important to triangulate one's publication bias assessments across levels of relevant study characteristics. Even assuming that a meta-analyst does assess bias across levels, multiple methods should be applied. Looking only at the trim-and-fill results, for instance, might give researchers the impression that publication bias does not vary across levels. The range of results overall, or even the Egger's regression results alone, can paint an entirely different picture. Thus, it appears that an assessment of publication bias at each level of relevant study characteristics may be crucial for understanding bias. Of course, asking meta-analysts to assess bias multiple times with multiple methods across multiple levels of study characteristics is demanding, especially considering that 53% of meta-analysts use only one method at all. A minimally acceptable goal may be for meta-analysts to consider their field of study, attempt to determine the most suspect study characteristic, and assess that variable. Not only will this help inspire meta-analysts to think more about potential sources of publication bias but doing so can reveal hidden sources of bias, or bias at one level that might not be found by assessing the overall data set.

Although, as demonstrated, one can determine whether patterns of bias differ by assessing publication bias multiple times across groups, there is a need for models of publication bias that can account for the role of study characteristics. We envision that such models would be able to accommodate differing publication bias patterns without allowing the mean model to differ across levels as well. This constraint on the mean model is necessary to distinguish between cases like Kling et al. (1999), where the mean effect does genuinely differ across levels and is apparently not an artifact of bias, and cases where the differing means can be better explained by different publication bias patterns. Trim and fill, modified to accommodate metaregression, may be able to consider both groups simultaneously and adjust for differing patterns of bias; although the results here are promising, nothing is known about this use of modified trim and fill, and an exploration of its performance through simulation is needed. Because there is evidence that study characteristics can affect publication bias, it would be useful to be able to model the impact of these variables, allowing meta-analysts to assess their influence and even adjust for it.

Future goals of this research include investigating more sophisticated methods for modeling the relationship between bias and study characteristics. One such method might involve adapting the weight-function model to estimate a parameter that can accommodate the differing weight patterns and provide adjusted estimates more economically than a model that estimates entirely separate sets of weights for the different study types. Modifying Egger's regression could also be plausible, perhaps by including a dummy-coded variable representing the suspect study characteristic, so that the interaction term could reflect bias in the second group. An ideal model would be able to estimate both categorical and continuous moderators, and would require the addition of minimal parameters to reduce the burden on estimation. Such a model could allow researchers to assess the degree of publication bias and its variation across study characteristics, and to examine whether such an issue is likely to be present in their data.

Conclusions

Assessing publication bias is a complicated issue. We recommend that meta-analysts give serious thought to the factors or study characteristics that might affect bias in their field; we also recommend that they consider coding for these factors, and comparing patterns of bias across levels. We would like to emphasize the importance of triangulation, or using multiple methods of assessment, and encourage meta-analysts to report and consider the results of several methods whenever possible. We have presented several approaches that could be used in that process, namely funnel plots, Egger's regression, trim and fill (with and without a metaregression algorithm), cumulative meta-analysis, and weight-function models. We encourage meta-analysts to think carefully about their data and assess patterns of publication bias across levels of viable study characteristics using as many of these approaches as seems applicable. Finally, we recommend that both researchers and consumers of research alike remember that publication bias is not always as simple as the suppression of nonsignificant results.

References

Abramowitz, S. I., Gomes, B., & Abramowitz, C. V. (1975). Publish or politic: Referee bias in manuscript review. *Journal of Applied Social Psychology, 5,* 187–200. http://dx.doi.org/10.1111/j.1559-1816.1975.tb00675.x

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Armstrong, J. S. (1996). Publication of research on controversial topics: The early acceptance procedure. *International Journal of Forecasting, 12,* 299–302. http://dx.doi.org/10.1016/0169-2070(95)00626-5

Banks, G. C., Kepes, S., & McDaniel, M. A. (2015). Publication bias: Understanding the myths concerning threats to the advancement of science. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 36–64). New York, NY: Routledge.

Barnes, D. E., & Bero, L. A. (1998). Why review articles on the health effects of passive smoking reach different conclusions. *Journal of the American Medical Association, 279,* 1566–1570. http://dx.doi.org/10.1001/jama.279.19.1566

Becker, B. J. (2005). Fail-safeN 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: Wiley.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society Series A, 151,* 419–463. http://dx.doi.org/10.2307/2982993

Begg, C. B., & Berlin, J. A. (1989). Publication bias and dissemination of clinical research. *Journal of the National Cancer Institute, 81,* 107–115. http://dx.doi.org/10.1093/jnci/81.2.107

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. http://dx.doi.org/10.2307/2533446

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *Journal of the American Medical Association, 289,* 454–465. http://dx.doi.org/10.1001/jama.289.4.454

Berlin, J. A., Begg, C. B., & Louis, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association, 84,* 381–392. http://dx.doi.org/10.1080/01621459.1989.10478782

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, United Kingdom: Wiley. http://dx.doi.org/10.1002/9780470743386

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,* 447–452. http://dx.doi.org/10.1037/1082-989X.2.4.447

Copas, J., & Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics, 1,* 247–262. http://dx.doi.org/10.1093/biostatistics/1.3.247

Dear, K. B. G., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science, 7,* 237–245. http://dx.doi.org/10.1214/ss/1177011363

Dickersin, K., & Min, Y. I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences, 703,* 135–148. http://dx.doi.org/10.1111/j.1749-6632.1993.tb26343.x

Doucet, M., & Sismondo, S. (2008). Evaluating solutions to sponsorship bias. *Journal of Medical Ethics, 34,* 627–630. http://dx.doi.org/10.1136/jme.2007.022467

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 127–144). Chichester, England: Wiley.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89–98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Duval, S., & Weinhandl, E. (2011). *Correcting for publication bias in the presence of covariates* (AHRQ Publication No. 11-EHC041-EF). Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from www.effectivehealthcare.ahrq.gov/reports/final.cfm

Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin, 129,* 569–591. http://dx.doi.org/10.1037/0033-2909.129.4.569

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991. April). Publication bias in clinical research. *Lancet, 337,* 867–872. http://dx.doi.org/10.1016/0140-6736(91)90201-Y

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629

Felson, D. T. (1992). Bias in meta-analytic research. *Journal of Clinical Epidemiology, 45,* 885–892. http://dx.doi.org/10.1016/0895-4356(92)90072-U

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research, 36,* 153–172. http://dx.doi.org/10.1177/0049124107306659

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research* (Vol. 56). Beverly Hills, CA: Sage.

Goodstein, L. D., & Brazis, K. L. (1970). Psychology of scientist: XXX. Credibility of psychologists: An empirical study. *Psychological Reports, 27,* 835–838. http://dx.doi.org/10.2466/pr0.1970.27.3.835

Heckman, J. J. (1995). Lessons from the bell curve. *Journal of Political Economy, 103,* 1091–1120. http://dx.doi.org/10.1086/262014

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics, 9,* 61–85. http://dx.doi.org/10.3102/10769986009001061

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7,* 246–255. http://dx.doi.org/10.1214/ss/1177011364

Herrnstein, R. J., & Murray, C. (1994). *Bell curve: Intelligence and class structure in American life*. New York, NY: Simon & Schuster.

Higgins, J. P. (Ed.). (2008). *Cochrane handbook for systematic reviews of interventions* (Vol. 5). Chichester, England: Wiley-Blackwell. http://dx.doi.org/10.1002/9780470712184

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104,* 53–69. http://dx.doi.org/10.1037/0033-2909.104.1.53

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science, 3,* 109–117. http://dx.doi.org/10.1214/ss/1177013012

Jennions, M. D., & Møller, A. P. (2002). Publication bias in ecology and evolution: An empirical assessment using the "trim and fill" method. *Biological Reviews of the Cambridge Philosophical Society, 77,* 211–222. http://dx.doi.org/10.1017/S1464793101005875

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24,* 602–611. http://dx.doi.org/10.2307/2392366

Johnson, D. H. (2001). Sharing data: It's time to end psychology's guild approach. *APS Observer, 14,* 38–39. Retrieved from http://www.psychologicalscience.org/observer/1001/data.html

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15,* 624–662. http://dx.doi.org/10.1177/1094428112452760

Kepes, S., Banks, G. C., & Oh, I. S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology, 29,* 183–203. http://dx.doi.org/10.1007/s10869-012-9279-0

King, G. (1995). Replication, replication. *PS: Political Science and Politics, 28,* 444–452.

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125,* 470–500. http://dx.doi.org/10.1037/0033-2909.125.4.470

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107–112. http://dx.doi.org/10.1111/j.2044-8317.1978.tb00578.x

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

Maccoby, E. E., & Jacklin, C. N. (Eds.). (1974). *The psychology of sex differences* (Vol. 1). Stanford, CA: Stanford University Press.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161–175. http://dx.doi.org/10.1007/BF01173636

McDaniel, M. A. (2009, April). *Cumulative meta-analysis as a publication bias method*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

McDaniel, M. A., McKay, P., & Rothstein, H. R. (2006, May). *Publication bias and racial effects on job performance: The elephant in the room*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX. http://dx.doi.org/10.1037/0021-9010.91.3.538

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59,* 927–953. http://dx.doi.org/10.1111/j.1744-6570.2006.00059.x

McKay, P. F., & McDaniel, M. A. (2006). A reexamination of black-white mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91,* 538–554. http://dx.doi.org/10.1037/0021-9010.91.3.538

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F., & The QUOROM Group. (1999, November). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *Lancet, 354,* 1896–1900. http://dx.doi.org/10.1016/S0140-6736(99)04149-5

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151,* 264–269, W64. http://dx.doi.org/10.7326/0003-4819-151-4-200908180-00135

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61,* 991–996. http://dx.doi.org/10.1016/j.jclinepi.2007.11.010

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Roseman, M., Milette, K., Bero, L. A., Coyne, J. C., Lexchin, J., Turner, E. H., & Thombs, B. D. (2011). Reporting of conflicts of interest in meta-analyses of trials of pharmacological treatments. *JAMA: Journal of the American Medical Association, 305,* 1008–1017. http://dx.doi.org/10.1001/jama.2011.257

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, England: Wiley. http://dx.doi.org/10.1002/0470870168

RStudio. (2012). *RStudio: Integrated development environment for R* (Version 0.96.122) [Computer software]. Boston, MA: Author. Available from http://www.rstudio.org/

Rufibach, K. (2011). Selection models with monotone weight functions in meta analysis. *Biom. J., 53,* 689–704.

Schneider, B. (2004). Building a scientific community: The need for replication. *Teachers College Record, 106,* 1471–1483. http://dx.doi.org/10.1111/j.1467-9620.2004.00386.x

Schwarzer, G. (2015). meta: Meta-analysis with R [Computer software]. R package, version 4.0–3.

Smith, M. L. (1980). Sex bias in counseling and psychotherapy. *Psychological Bulletin, 87,* 392–407. http://dx.doi.org/10.1037/0033-2909.87.2.392

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5,* 60–78. http://dx.doi.org/10.1002/jrsm.1095

Sterne, J. A., Becker, B. J., & Egger, M. (2005). The funnel plot. *Publication bias in meta-analysis: Prevention, assessment, and adjustments,* 75–98. http://dx.doi.org/10.1002/0470870168.ch5

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54,* 1046–1055. http://dx.doi.org/10.1016/S0895-4356(01)00377-8

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal, 323,* 101–105. http://dx.doi.org/10.1136/bmj.323.7304.101

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113–2126. http://dx.doi.org/10.1002/sim.1461

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. http://dx.doi.org/10.1007/BF02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10,* 428–443. http://dx.doi.org/10.1037/1082-989X.10.4.428

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

Weinhandl, E. D., & Duval, S. (2012). Generalization of trim and fill for application in meta-regression. *Research Synthesis Methods, 3,* 51–67. http://dx.doi.org/10.1002/jrsm.1042

Received September 30, 2014
Revision received May 15, 2015
Accepted May 20, 2015 ■