*Communication Monographs*
*Vol. 76, No. 3, September 2009, pp. 286–302*

Routledge
Taylor & Francis Group

# Sample Sizes and Effect Sizes are Negatively Correlated in Meta-Analyses: Evidence and Implications of a Publication Bias Against NonSignificant Findings

Tim Levine, Kelli J. Asada & Chris Carpenter



*Meta-analysis involves cumulating effects across studies in order to qualitatively summarize existing literatures. A recent finding suggests that the effect sizes reported in meta-analyses may be negatively correlated with study sample sizes. This prediction was tested with a sample of 51 published meta-analyses summarizing the results of 3,602 individual studies. The correlation between effect size and sample size was negative in almost 80 percent of the meta-analyses examined, and the negative correlation was not limited to a particular type of research or substantive area. This result most likely stems from a bias against publishing findings that are not statistically significant. The primary implication is that meta-analyses may systematically overestimate population effect sizes. It is recommended that researchers routinely examine the* n–r *scatter plot and correlation, or some other indication of publication bias and report this information in meta-analyses.*

*Keywords: Meta-Analysis; Effect Size; Publication Bias*

> Those familiar with human nature and the publication process acknowledge that biases due to selective reporting of results are likely widespread in all fields of academic inquiry that depend on tools of statistical inference.
>
> —A. Richard Palmer (1999, p. 220)

Timothy R. Levine is Professor of Communication at Michigan State University where Chris Carpenter is a doctoral student. Kelli J. Asada is Assistant Professor of Communication at William Paterson University. Correspondence to: Tim Levine, Michigan State University, Communication, 482 CAS, East Lansing, MI 48824, USA. E-mail: levinet@msu.edu

ISSN 0363-7751 (print)/ISSN 1479-5787 (online) 
DOI: 10.1080/03637750903074685

> Statistical combination of studies from the literature can be trusted to be unbiased only if there is reason to believe that there are essentially no unpublished studies [which is] almost never the case!
>
> —Jeffery D. Scargle (1999, p. 22)

Meta-analysis is an increasing popular and influential way of summarizing the results of quantitative research. One advantage of meta-analyses is that it focuses attention on effect sizes rather than null hypothesis significance testing. One advantage of attention to effect sizes over significance tests, in turn, is that they are, in principle, independent of sample size. A recent finding (La France, Heisel, & Beatty, 2004), however, suggests that in actual practice, this may not be the case. Instead, sample sizes and effect sizes may often be negatively correlated in existing communication and other social and life science literatures. Many may find such a claim surprising and counterintuitive. Readers may wonder what this means and why they should be concerned.

Documenting a negative correlation between sample size and effect size would have several important implications. Such a finding challenges the prevailing belief that the two are independent. It may suggest that meta-analyses tend to overestimate effect sizes. At the very least, if such a correlation is prevalent, meta-analysts need to routinely check for this association, and report the *n*–*r* correlation as standard practice. Solutions would need to be devised and implemented. All this, however, first requires documentation that the problem actually exists.

This paper reports a meta-analysis of meta-analyses to assess if sample size is empirically related to effect size across a variety of research literatures related to communication. Effect size and sample size should be relatively independent (Hunter & Schmidt, 1990; Levine, Weber, Hullett, Park, & Lindsey, 2008). If such a correlation exists in some substantial proportion of published meta-analyses, this would suggest a pervasive bias or artifact in the results of meta-analysis. Possible reasons for such a bias are considered along with tentative solutions.

## Meta-Analysis

Meta-analysis is a study of studies. Specifically, it is a set of methods and statistical analyses for summarizing the findings of existing quantitative, empirical literatures. Meta-analysis assesses if the findings from a collection of studies investigating some specific issue lead to some consistent result and, if so, the magnitude of that finding is estimated. This is done by cumulating quantitative effects across studies.

Meta-analysis involves several steps. First, relevant and usable studies investigating a topic are collected. Then, the findings of each study is converted to some common metric so that the results can be cumulated. Relevant study features are also coded. Then, an average effect across studies is calculated, and study-to-study variability is examined. Analyses are also done to see if and how coded study features affect results.

Once previous studies have been collected, the findings from each study need to be converted to a common metric, usually some unit of "effect size." The most common metrics used in meta-analysis are *d* and *r*. *d* is the standardized mean difference, and *r* is the correlation coefficient. So, for each test of a hypothesis in the literature, an

effect size is obtained. If the previous studies report effect sizes, this is straight forward. If effect sizes are not reported, a variety of conversion formulas exist. For example, if either sufficient descriptive statistics (e.g., means, standard deviations, and cell sizes) or significance tests with degrees of freedom are reported, effect sizes can usually be calculated.

Once a set of effects has been collected reflecting the findings in the literature, the findings are cumulated and tested for homogeneity of effects. Findings are cumulated simply by averaging, although the average is usually weighted by study sample. This produces an across-study average effect, and this average effect can be considered an estimate of the population effect. The across-study average can be tested to see if it is likely different than zero, and confidence intervals can be calculated around the average. Because across-study average effects are based on much larger and more diverse samples than any single study, meta-analysis provides a better and more stable picture of the state of a research literature than is obtained from a less systematic examination of individual studies.



## Significance, Effect Size, and Sample Size

Tests of statistical significance tests are typically employed as the primary decision strategy in most quantitative communication research. Although the problems with significance testing are many, perhaps the most widely recognized limitation in statistical significance testing is its sensitivity to sample size (e.g., Boster, 2002; Levine, Weber, Park, et al., 2008). When the sample size is small, strong and important effects can be nonsignificant, but when sample sizes are large, even trivial effects can have impressive looking *p*-values. Thus, the *p*-values from null hypothesis significance tests reflect both the sample size and the magnitude of the effect observed, and obtaining or failing to obtain statistical significance is as much a function of one's sample size (and other things that affect statistical power like measurement reliability, manipulation strength, meeting statistical assumptions etc.) as the verisimilitude of one's substantive hypothesis (Meehl, 1986).

As a consequence of the sample size problem, there is a growing recognition of the importance of reporting and interpreting effects sizes to supplement significance tests (Levine, Weber, Park, et al., 2008). Various estimates of magnitude of effect or effect size tell us how strongly two or more variables are related, or how large is the difference between groups. Theoretical and practical importance rest more on the magnitude of effect than on the probability of the data given the null hypothesis (Abelson, 1995; Boster, 2002; Cohen, 1994).

One virtue of meta-analysis is its focus on effect size (Boster, 2002). Meta-analyses provide estimates of population effect sizes. In doing so, it avoids some pitfalls associated with the use of significance tests in single studies (Levine, Weber, Hullett, et al., 2008). The large sample sizes obtained from cumulating results across studies lead to increased statistical power, narrower of confidence intervals, and stronger evidence for the generality of a finding (Abelson, 1995).

In principle, sample size and effect size should be unrelated. How strongly two variable covary, or the extent to which a set of means are different from one another should not be a function of the mere number subjects used in the test. Instead, estimates of effect size should be relatively unbiased estimates of population effects.

While effect size is independent of sample size in principle, this may not be the case in practice. La France et al., (2004) report a meta-analysis of extroversion and nonverbal behavior and observed a substantial negative correlation ($r = -.40$, $p < .10$) between estimates of effect size and the sample size in the individual studies included in their meta-analysis. One might wonder if La France et al.'s results are typical of meta-analyses or just an anomaly.

On one hand, Hunter and Schmidt (1990) claim that a correlation between sample size and effect size is highly unusual, and explainable by atypical methodological artifact. If Hunter and Schmidt are correct, then the average correlation between effect size and sample size should approach zero. Further, while very few correlations would be expected to be exactly zero, the frequency of positive and negative correlations should be approximately equal. With the meta-analytic test as the unit analysis, one could dichotomize the observed *n*–*r* correlations as positive or negative and expect a binomial distribution centered on an outcome probability of $p = .50$.

On the other, La France et al. (2004) reviewed three recent meta-analyses in addition to their own and report a negative correlation in each (average $r = -.35$). Four relatively small meta-analyses, however, may not be especially representative, and consequently they provide only limited evidence for a wide-spread phenomenon. Nevertheless, their findings suggest that negative *n*–*r* correlations occur. To gage the approximate frequency of this finding, a first research question is posed.

RQ1: Approximately how prevalent is the correlation between sample size and effect size in published meta-analyses and is the frequency greater than that expected by chance?

## Cognitive Overload

La France et al. (2004) explain their negative *n*–*r* correlation with a cognitive overload account. The dependent measures in the studies included in their meta-analysis were coded nonverbal behaviors. Larger sample sizes meant more work for coders, and overworked coders might be less reliable. Lower reliability would result in smaller effect sizes. Thus, larger studies might find smaller effects as a consequence of differential coding quality. Consistent with this, they also observed an association between effect sizes and the number of behaviors coded. The fewer behaviors coded in a given study, then larger the effects. Thus, in some literatures, sample size may be proxy for research quality, with higher quality studies finding larger effects.

Although La France et al.'s (2004) explanation is plausible and likely has some merit, a more parsimonious and robust (but not mutually exclusive) explanation is possible. A negative *n*–*r* correlation between sample size and effect size in meta-analysis would be anticipated if there existed a preference for statistically significant

findings. That is, publication bias can predict and explain a negative association between effect sizes and sample sizes in published research.

## Publication Bias

There is a growing consensus over the past decade outside the field of communication that publication bias is wide spread and pervasive (Gerber & Malhotra, 2008; Thornton & Lee, 2000). "A publication bias exists if the probability that a study reaches the literature, and is thus available for combined analysis, depends on the results of the study" (Scargle, 1999, p. 6). It has been shown that publication bias can substantially impact meta-analytic results to the point of qualitatively altering the conclusions drawn from meta-analysis (Palmer, 1999; Sutton, Duval, Tweedie, Abrams, & Jones, 2000).

Social and life scientists, including communication researchers, rely heavily on null hypothesis significance tests, and tend to conflate statistical significance with substantive importance (Boster, 2002; Levine, Weber, Park, et al., 2008). As a consequence, authors are less likely to submit research yielding nonsignificant findings for peer review, and journal reviewers and editors are less likely to publish nonsignificant findings (Callaham, Wears, Weber, Barton, & Young, 2008; Gerber & Malhotra, 2008; Meehl, 1986). Published work, in turn, is more likely to be included in subsequent meta-analysis (Kromrey & Rendina-Gobioff, 2006). Thus, achievement of $p < .05$ is a likely predictor of publication and subsequent inclusion in meta-analysis (Littner, Mimouni, Dollberg, & Mandel, 2005). Achieving $p < .05$, in turn, is a strong function of sample size.

Small-sample studies can only obtain statistically significant results when the effect sizes are substantial. Further, studies with small samples produce less stable results. Across small sample studies, isolated large effects can be obtained by chance. So, as sample size decreases, there will be more study-to-study variability, with the relatively larger findings being more likely to make it into print, and hence into meta-analyses.

As sample sizes increase, so does statistical power, and smaller effects can be $p < .05$. At the same time, there will be less across-study variability, and the effect size estimates will fall closer to the population values. Because the most population effect sizes in social science meta-analyses are moderate to small (Richard, Bond & Stokes-Zoota, 2003), large studies can and do find such effects. Smaller studies, however, report, on average, larger effects because smaller effects are nonsignificant and don't make it into print, and because larger effects occur more often (because of wider confidence intervals). The net result is a negative *n*–*r* correlation.

Thinking along the same lines, Kromrey and Rendina-Gobioff (2006) made the following argument:

> When the decision to publish is based on statistical significance, there is a direct relationship between publication probability and sample size. Essentially, the greater the sample size, the greater the chances of a statistically significant result and thus a greater chance of being published. Research with small samples and large treatment effects are more likely to be published than are small sample studies with

> smaller treatment effects. This results in a relationship between the effect sizes and sample sizes in the published literature. (p. 358)

If this reasoning is correct, a negative *n–r* correlation often exists and stems from a systemic bias in social and life science research making the bias both prevalent and not tied a specific method (e.g., coded dependent measures), topic, or field. This alternative explanation leads to a second research question. La France et al. suggest that the *n–r* correlation may be limited to studies with coded dependent measures, where as a publication bias would be more general, and transcend both topic and method.

> RQ2: Does research method moderate the correlation between sample size and effect size?

Given the existence of publication bias has been recognized at least since the mid 1950s (Thornton & Lee, 2000; Dickersin, 2005, however traces concerns back to the late 1700s), it is not surprising that several methods of testing for publication bias have been developed. The oldest and most well known is Rosenthal's (1979) failsafe *N*. The failsafe *N* is used to calculate the number of unpublished studies with an effect size of zero that would have to exist in order for the average effect size estimated by the meta-analysis to be no longer statistically significant (Rosenthal, 1979). However, there are many problems with failsafe *N* including widespread misinterpretation of the result (Becker, 2005), reliance on unrealistic assumptions (Scargle, 1999) and its failure to account for studies that were suppressed due to significant findings in the opposite direction (Begg, 1994). Failsafe *N* typically underestimates the extent of bias considerably (Scargle, 1999).

An alternative method of detecting publication bias that is gaining popularity is the examination of the funnel plot (Light & Pillemer, 1984). The funnel plot is a scatter plot that presents the effect sizes of primary studies included in a meta-analysis on the *x*-axis and the sample size of the study on the *y*-axis (see Sterne, Becker, & Egger, 2005 for a discussion of other choices for the *x* and *y* axes). Based on the assumption that studies with smaller samples will vary more around the true effect size than larger studies, a funnel plot with no publication bias will form a funnel shape. If there are few data points to the left of the true effect size, the funnel will appear to have a piece missing. Light and Pillemer argue this may be due the absence of nonsignificant findings or significant findings in the opposite direction expected in the literature base. They also note that funnel plot asymmetry may also be caused by the effect sizes coming from more than one population. A significance test based on the rank correlation between effect sizes and variances is provided by Begg and Mazumdar (1994), but the Begg rank correlation test is known to be underpowered (Duval & Tweedie, 2000b).

The funnel plot also faces problems. Vevea and Woods (2005) demonstrated that there are situations where visual examination of the funnel plot may make it difficult to detect evidence of subtle but important publication bias. Examination of a funnel plot has also been criticized as highly subjective and therefore difficult to assess consistently (Rothstein, 2008). To overcome these objections, new methods have been

developed to test for funnel plot asymmetry. The most promising may be the trim and fill (Peters, Sutton, Jones, Abrams, & Rushton, 2007).

The trim and fill (Duval & Tweedie, 2000a, 2000b) was developed to estimate the number of missing studies, add them to the distribution to achieve symmetry, and then recalculate the mean effect size. The method offers several computationally simple statistics to estimate the number studies missing from the left side of the funnel plot. This number of studies is then removed from the right side of the funnel. The mean effect size is then recalculated. The number of studies missing from the left size is again calculated and an equal number are again removed from the right side. Iterations of this process are conducted until the mean effect size stabilizes. Duval (2005) has found that this usually only takes two or three iterations. The studies that were removed are then put back. The studies that are thought to be missing are then filled in by adding studies to the left of the mean effect size that are symmetrical to the studies that were removed from the right side during the iterations. Then the mean effect size is again calculated and the new estimate of the mean effect size should be compared to the original mean effect size. If the trim and fill estimate of the mean effect is discrepant from the original, publication bias is indicated. Unfortunately, like the funnel plots, the trim and fill can indicate publication bias due to heterogeneity of effects where no publication bias is present (Peters et al., 2007; Terrin, Schmid, Lau, & Olkin, 2003).

The existence of publication bias has become generally accepted in the social and life sciences. A number of strategies currently exist for detecting potential publication biases in meta-analysis. These observations lead to a final three part research question.

RQ3: Are methods of detecting publication typically reported in meta-analysis, and if so, which methods are most common, and how common are they?

## Method

The data for the present study were taken from a sample of published meta-analyses. The only formal criterion for inclusion was that the meta-analysis report effect sizes and sample sizes for the individual studies that were included in the meta-analysis. This criterion was necessary in order to test the current research questions. Meta-analyses were collected with a target of obtaining 50 meta-analyses meeting the inclusion criteria. A total of 51 meta-analyses provided the data for the present analyses. These 51 meta-analyses summarized the results of 3,602 individual studies and yielded 75 *n*–*r* correlations.

Efforts were made to cover a wide range of topics relevant to communication researchers. Initially, meta-analyses published in Allen's two collections of meta-analyses on persuasion and interpersonal communication were included (Allen & Preiss, 1998; Allen, Preiss, Gayle, & Burrell, 2002). Meta-analyses were also obtained from a variety of journals (e.g., *CM, CR, HCR, JOC, JPSP*) and summarized a variety of literatures (e.g., attitude–behavior relationship, exposure to pornography, language

intensity effects, self-construals and culture, viewing presidential debates, violent video games). A complete list of the meta-analyses included is provided in Appendix A.

The unit of analysis was the 75 population effects tested. Some meta-analyses tested more than one effect, and consequently contributed multiple data points. The effect size metric used in the current analysis was *r*. For meta-analyses reporting effect sizes in *d*, *d* was first converted to *r*. A correlation was then calculated between primary study sample size ($n$) and the absolute value of the reported effect size /*r*/ for each population effect tested. The number of individual studies included in the 51 meta-analyses, and the sample sizes for the *n*–*r* correlations, ranged from 7 to 436 ($M = 48.03$, $SD = 64.95$). Each meta-analysis was coded for predominant dependent measure type, design type, and the mention of publication bias. Dependent measures were classified as coded ($n = 7$), scaled ($n = 46$), or mixed ($n = 22$). Designs were coded as experimental ($n = 36$), self-report/survey ($n = 33$) or mixed ($n = 6$). The mean sample size of the primary studies, the mean effect size of the primary studies, and the number of primary studies was also recorded.

Initial coding was done by the second author. The first author independently coded 10 meta-analyses, and the intercoder agreement was 100 percent (kappa = 1.00).



## Results

The *n*–*r* correlations ranged from $-.83$ to $+.60$. The mean correlation was $-.16$. Of the 75 correlations examined, 14 (18.6 percent) were positive, 2 (2.6 percent) rounded to zero, and 59 (78.6 percent) were negative. A binomial test contrasting the observed frequency of negatively signed correlations against a 50 percent base-rate suggested that the observed frequency of negatively signed correlations was not attributable to chance at $p < .00001$. An on-line calculator was used for this and subsequent binomial tests (http://www.stat.tamu.edu/ ~west/applets/binomialdemo.html).

Further, 16 of the 75 correlations were statistically significant at $p < .05$, a number of correlations that is significantly ($p < .00001$) greater than that expected by chance if the null hypothesis were true. Of these 16, 14 were negative and two were positive. Obtaining 14 significant negative correlations at $p < .05$ by chance if the null was always true would be highly improbable, $p < .00001$. Alternatively, obtaining at least two statistically significant positive correlations by chance is not surprising under the null, $p = .56$. Thus, the data clearly show that the correlation between sample size and effect size tends to be negative. Although the mean correlation is not large, it is more consistently negative than an independent *n*–*r* model allows.

Analysis by dependent measure type did not yield evidence of differential results. Coded (mean $r = -.16$), scaled ($r = -.14$) and mixed ($r = -.20$) literatures all yielded negative *n*–*r* correlations, and the difference was trivial, $F(2,72) < 1.00$, $p = .75$. For design type, the *n*–*r* correlation was somewhat stronger for experimental studies ($-.21$) than for self-report literatures ($-.11$), but the difference was not statistically significant, $p = .12$. Experimental studies, however, had much smaller sample sizes on average than survey studies (experimental average $n = 118$,

nonexperimental average $n = 248$, $t = 3.16$, $p < .001$). The marginal effect for design type on the *n–r* correlation disappeared when controlling for the average sample size in the primary studies, $F < 1.00$, $p = .63$. Thus, the finding seems to be general across design types and dependent measurement types and the apparent differences were an artifact of differential sample sizes in the primary studies.

Other moderators that were examined included the average sample size of the primary studies included in the meta-analysis, the average effect size reported by the meta-analysis, and the number of primary studies included in the meta-analysis. Zero-order correlations between the *n–r* correlation and these study features were calculated. The correlations were $r = .19$, $p = .098$ for mean sample size, $r = .05$, $p = .67$ for absolute mean effect size, and $r = .08$, $p = .50$ for number of studies. Examining only those studies with negative *n–r* correlations, medium and large negative *n–r* correlations occurred predominantly in literatures where the average small size was less than 200, where the average effect size was less than $r = .40$, and in meta-analysis containing less than 100 individual studies. Moderated regression analysis failed to provide evidence of statistical interactions among these factors.

Qualitative examination of the studies for trends reveals the statistically obvious. A plot of the *n–r* correlations reveals a symmetrical, unimodal distribution centered around the mean correlation (see Table 1). Just as one might expect, meta-analyses yielding *n–r* correlations widely discrepant from the average correlation tend to be based on a relatively small number studies. For, example, the meta-analysis yielding the +.60 correlation involved 13 studies and the −.83 correlation involved only 9

**Table 1** Stem and Leaf Plot of the Correlation Between Study Sample Size and Reported Effect Size in Published Meta-Analyses

| | |
|---|---|
| −.8 | **3** |
| −.7 | |
| −.6 | **43** |
| −.5 | 3 |
| −.4 | 96**65**5**3**2**1** |
| −.3 | 86**5**300**00** |
| −.2 | **9**99776554330 |
| −.1 | 9**88**8665442100 |
| −.0 | 888766553221 |
| +.0 | 00124668 |
| +.1 | |
| +.2 | 0**3**49 |
| +.3 | 9 |
| +.4 | 7 |
| +.5 | 5 |
| +.6 | **0** |
| +.7 | |
| +.8 | |

*Note.* Mean correlation = −.16 (SE = .03), median −.18, 79% of correlations are negative. Mean $K = 48$ (range 7 to 436), correlation between $K$ and $r = .08$, $p = ns$. Bold in table indicates correlations that are significant at $p < .05$

studies. Alternatively, the large meta-analyses reviewing numerous studies consistently show small to moderate and often statistically significant negative *n–r* correlations.

The *n–r* relationship within each meta-analysis was examined with a series of scatter plots. Qualitative examination suggests that the negative *n–r* relationship is most often apparent in literatures (and meta-analyses) characterized by substantial variance in the sample sizes of individual studies, and small to moderate effects average effects for studies with larger samples. A scatter plot of the studies included in Allen et al. (1989) provides an example (see Figure 1). The mean weighted effect size is $r = .29$. The average absolute effect for studies with $N < 100$ is $r = .35$ while studies with $N > 100$ have a mean effect of $r = .18$, a difference that is statistically significant, $t\ (434) = 2.74$, $p < .01$. Further the nonlinear negative but decelerating function was typical, and consistent with that observed by La France et al. (2004).

The 51 meta-analyses sampled were screened for mention of negative *n–r* correlations or mention of publication bias. Only 8 of the 51 meta-analysis (16 percent) addressed these issues in any manner. Three studies mentioned publication bias in text but did not report a test. Three studies reported fail-safe *N*. Only a single study reported a funnel plot. Only La France et al. (2004) reported a *n–r* correlation.



## Discussion

The current investigation examined the association between sample size (*n*) and effect size (*r*) in 3,602 studies included in 51 published meta-analyses. In principle, sample

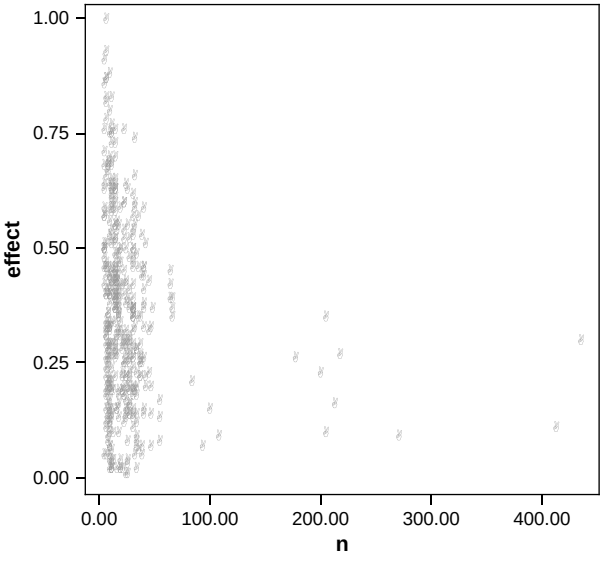


**Figure 1** Example scatter plot of sample sizes and effect sizes in a meta-analysis. Data were taken from Allen et al. (1989). The *n–r* correlation is $r\ (434) = -.179$, $p < .001$. Mean weighted effect $r = .288$. Mean weighted effect for studies with $N < 100$, $r = .318$. Mean weighted effect for studies with $N > 100$, $r = .187$.

size and effect size are (mostly) statistically independent. In practice, however, they are not. The *n–r* correlation was negative in nearly 80 percent of the meta-analyses examined. The predominately negative association between sample size and effect size observed here is systematic and almost certainly not attributable to chance. Something systematic is operating.

Because the negative *n–r* correlation was not confined to a particular literature, research design, or type of dependent measure (e.g., nonverbal coding), the finding is general and likely stems from some pervasive process or phenomena. The current authors believe the most plausible explanation is a bias against nonsignificant findings. If researchers are less likely to make their nonsignificant results public, and if journal reviewers and editors are more likely to accept statistically significant rather than nonsignificant results, we might expect a negative *n–r* correlation. Large sample studies produce more stable population estimates and have the statistical power necessary to detect more meager effects. Smaller studies produce a wider range of findings, but only the relatively large findings get reported because those are the ones that are statistically significant. Hence, the negative correlation.

Concern of publication bias is evident in most of the social and life sciences including biology, education, epidemiology, political science, and sociology (see Gerber & Malhotra, 2008 for a recent review). Interestingly, however, this concern was evident in only a relatively few of the meta-analyses sampled in this investigation. Because negative *n–r* correlations were prevalent in the meta-analysis sampled, this lack of awareness appears problematic.

If the reasoning presented here is correct, the primary substantive implication of these findings is that meta-analyses systematically present an overly rosy depiction of the literature. Given that the average effect size reported in meta-analyses in the social sciences are meager to begin with (Richard et al., 2003), if an upward bias exists, most effects may be even smaller that than current believed.

The case of the meta-analysis by Allen et al. (1989; see Figure 1) illustrates this point nicely. The *n–r* correlation in Allen et al. (1989) was $-.179$, a value close the across-meta-analysis average of $-.16$. The mean weighted effect observed in the meta-analysis was $r = .288$. However, considering only studies with $N > 100$, the mean weighted effect drops to $r = .187$, a reduction of 35 percent. To the extent that the larger sample studies better estimate true population effects, the efficacy of treatments of public speaking anxiety were substantially over estimated.

Worse still, the current analysis may underestimate the magnitude of the problem. The sample-size-to-sampling-error relationship and power curves are nonlinear. Consistent with this, La France et al. (2004) show a nonlinear pattern. The curve was sharply negative at relatively small sample sizes and flattened out for larger studies. The current analysis, however, only estimated the linear association between sample size and effect size. Because the regression of sample size onto effect size should be nonlinear, many negative correlations may be underestimated, and hence the mean correlation was depressed.

The current findings do not mean that the La France et al.'s (2004) cognitive load hypothesis is false or discredited. Sample size can be a marker for study quality, and

study quality can be systematically related to the size of effects observed. There is nothing wrong with the logic of the La France et al. argument. But, whereas variability in study quality is probably sufficient to produce a *n–r* correlation, it is probably not the only culprit. Further, variable study quality could conceivably produce positive *n–r* correlations. It is not difficult to imagine literatures in which the studies with larger samples also tend to be those with tighter designs and better measurement. Thus, to the extent that the *n–r* association is a function of differential research quality, studies coding nonverbal behaviors might exhibit a different *n–r* pattern than, for example, program evaluation research.

In the short term, researchers conducting meta-analyses need to examine and report the *n–r* relationship, a funnel plot, trim and fill, or similar check in their analyses. If the *n–r* correlation is examined, both correlations and scatter plots need to be examined because nonlinearity might be anticipated. Whereas the *n–r* correlation should not substitute for funnel plots or other more sophisticated tests of publication bias, the *n–r* correlation is a simple, objective, and easy test that may provide a rough hint that bias exists in a literature. Further, the *n–r* correlation may result in fewer type-one errors than alternative methods of detecting publication bias (Kromrey & Rendina-Gobioff, 2006). Thus, if the *n–r* correlation is statistically significant, bias likely exists. Unfortunately, however, all the methods for detecting for detecting publications reviewed here appeared to be underpowered (Kromrey & Rendina-Gobioff, 2006; Masaskill, Walter, & Irwig, 2001). Thus, a lack of a significant correlation does not mean that no bias exists.

When bias is discovered or suspected, simply knowing that it exists can prompt a search for an explanation and a need to qualify conclusions accordingly. Those looking for a correction method might consider the trim and fill method (Duval & Tweedie, 2000a, 2000b). Although Duval and Tweedie do endorse using trim and fill as a correction method, it could be used as such so long as findings are homogeneous.

Ideally, however, social change directed at the root of the problem is needed. Whereas reliance on null hypothesis significance testing is not a likely candidate for extinction, authors, reviewers, and editors could and should be more open to considering the value of nonsignificant results. Granted, at the level of the individual study, a nonsignificant finding does not allow for much in the way of substantive conclusions. After all, we typically do not accept the null in traditional significance testing. Rejecting research on this basis alone, however, is a short-sighted view with negative consequences. Nonsignificant findings add to the larger literature, especially with the advent of meta-analysis. If only supportive findings are made public, a biased picture of the literature inevitably results. Better conclusion can be drawn when more information that is less biased information is available.

A potential concern with the current analysis is the relatively haphazard sampling of meta-analyses. Whereas it might have been possible to randomly select from a larger pool of meta-analyses, it was decided that the current efforts were better spent including all meta-analyses obtained in the analyses. While some inadvertent bias in study selection is possible, the current results are strong and it is difficult to imagine how an inadvertent selection bias could explain the results. Similarly, the study

coding was less formal the many would like. Nevertheless, it is unlikely that the coding was flawed to the extent that would alter the results in some dramatic way. If there was some strong but unspecified or mis-specified moderator or moderators operating, then a bi-, multimodal, or relatively flat distribution would be expected. As the stem-and-leaf plot shows, however, this is not the case. Instead, the *n*–*r* correlations exhibit a uni-modal, symmetrical distribution that lends confidence to the current findings and conclusions.

In conclusion, the current analysis finds strong evidence of a negative correlation between sample size and effect size in studies that get included in meta-analyses. This most likely stems from a bias against nonsignificant findings, and it likely results in meta-analyses overestimating effect sizes. Researchers doing meta-analyses need to check for and report the association between *n* and *r*, and research consumers need to exhibit caution in interpreting results. These findings also provide evidence consistent with arguments calling for changes in statistical practice (e.g., Cohen, 1994; Boster, 2002; Levine, Weber, Park, & Hullet, 2008; Meehl, 1986).



## References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: LEA.

Allen, M., & Preiss, R. W. (1998). *Persuasion: Advances through meta-analysis.* Cresskill, NJ: Hampton Press.

Allen, M., Hunter, J. E., & Donohue, W. A. (1989). Meta-analysis of self-report data on the effectiveness of public speaking anxiety treatment. *Communication Education*, *38*, 54–76.

Allen, M., Preiss, R. W., Gayle, B. M., & Burrell, N. A. (2002). *Interpersonal communication research: Advances through meta-analysis* (pp. 263–279). Mahwah, NJ: LEA.

Becker, B. J. (2005). Failsafe *N* or file-drawer number. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments* (pp. 111–125). Chichester, UK: Wiley.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 399–409). New York: Russell Sage Foundation.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101.

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research*, *28*, 473–490.

Callaham, M. L., Wears, R. L., Weber, E. J., Barton, C., & Young, G. (2008). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *Journal of the American Medical Association*, *280*, 254–257.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments* (pp. 12–33). Chichester, UK: Wiley..

Duval, S. (2005). The trim and fill method. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments* (pp. 127–144). Chichester, UK: Wiley.

Duval, S., & Tweedie, R. L. (2000a). A non-parametric "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. L. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 276–284.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and Research*, *37*, 3–30.

Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Kromrey, D. J., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement*, *66*, 357–373.

La France, B. H., Heisel, A. D., & Beatty, M. J. (2004). Is there empirical evidence for a nonverbal profile of extraversion?: A meta-analysis and critique of the literature. *Communication Monographs*, *71*, 28–48.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*, 188–209.

Levine, T. R., Weber, R., Hullett, C. R., Park, H. S., & Lindsey, L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*, 171–187.

Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Littner, Y., Mimouni, F. B., Dollberg, S., & Mandel, D. (2005). Negative results and impact factor: A lesson from neonatology. *Archives of Pediatric and Adolescent Medicine*, *159*, 1036–1037.

Masaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*, 641–654.

Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Meta-Theory in Social Science* (pp. 315–338). Chicago: University of Chicago Press.

Palmer, A. R. (1999). Detecting publication bias in meta-analysis: A case study of fluctuating asymmetry and sexual selection. *American Naturalist*, *154*, 220–233.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*, 4544–4562.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

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

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology*, *4*, 61–81.

Scargle, J. D. (1999). *The "file-drawer" problem in scientific inference*. Paper presented at the Sturrock Symposium. Stanford University.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments* (pp. 75–98). Chichester, UK: Wiley.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analysis. *British Medical Journal*, *320*, 1574–1577.

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.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*, 207–216.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*, 428–443.

## **Appendix A** Meta-Analyses Included as Data in Current Study

Allen, M., Bourhis, J., Emmers-Sommer, T., & Sahlstein, E. (1998). Reducing dating anxiety: A meta-analysis. *Communication Reports*, *11*, 49–55.

Allen, M., D'Alessio, D., & Brezgel, K. (1995). A meta-analysis summarizing the effects of pornography II: Aggression after exposure. *Human Communication Research*, *22*, 258–283.

Allen, M., D'Alessio, D., Emmers, T. M., & Gebhardt, L. (1996). The role of educational briefings in mitigating effects of experimental exposure to violent sexually explicit material: A meta-analysis. *The Journal of Sex Research*, *33*, 135–141.

Allen, M., Emmers-Sommer, T. M., & Crowell, T. L. (2002). Couples negotiating safer sex behaviors: A meta-analysis of the impact of conversation and gender. In M. Allen, R. W. Preiss, B. M. Gayle, & N. A. Burrell (Eds.), *Interpersonal communication research: Advances through meta-analysis* (pp. 263–279). Mahwah, NJ: Lawrence Erlbaum Associates.

Allen, M., Hunter, J. E., & Donohue, W. A. (1989). Meta-analysis of self-report data on the effectiveness of public speaking anxiety treatment. *Communication Education*, *38*, 54–76.

Allen, M., & Preiss, R. W. (1997). Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. *Communication Research Reports*, *14*, 125–131.

Beck, R., & Fernandez, E. (1998). Cognitive-behavioral therapy in the treatment of anger: A meta-analysis. *Cognitive Therapy and Research*, *23*, 63–74.

Benoit, W. L. (1998). Forewarning and persuasion. In R. W. Allen & Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 139–154). Cresskill, NJ: Hampton Press.

Benoit, W. L., Hansen, G. J., & Verser, R. M. (2003). A meta-analysis of the effects of viewing US presidential debates. *Communication Monographs*, *70*, 335–350.

Bettencourt, B. A., & Kernahan, C. (1997). A meta-analysis of aggression in the presence of violent cues: Effects of gender differences and aversive provocation. *Aggressive Behavior*, *23*, 447–456.

Burrell, N. A., & Koper, R. J. (1998). The efficacy of powerful/powerless language on attitudes and source credibility. In R. W. Allen & Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 203–215). Cresskill, NJ: Hampton Press.

Cohn, L. D., & Westenberg, P. M. (2004). Intelligence and maturity: Meta-analytic evidence for the incremental and discriminant validity of Loevinger's measure of ego development. *Journal of Personality and Social Psychology*, *86*, 760–772.

Cruz, M. G. (1998). Explicit and implicit conclusions in persuasive messages. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 217–230). Cresskill, NJ: Hampton Press.

DeCoster, J., & Claypool, H. M. (2004). A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Personality and Social Psychology Review*, *8*, 2–27.

Dillard, J. P., Hunter, J. E., & Burgoon, M. (1984). Sequential-request persuasive strategies: Meta-analysis of foot-in-the-door and door-in-the-face. *Human Communication Research*, *10*, 461–488.

Emmers-Sommer, T. M., Allen, M., Bourhis, J., Sahlstein, E., Laskowski, K., Falato, W. L., et al. (2004). A meta-analysis of the relationship between social skills and sexual offenders. *Communication Reports*, *17*, 1–10.

Fejfar, M. C., & Holye, R. H. (2000). Effect of private self-awareness on negative affect and self-referent attribution: A quantitative review. *Personality and Social Psychology Review*, *4*, 132–142.

Gayle, B. M., Preiss, R. W., & Allen, M. (1998). Another look at the use of rhetorical questions. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 189–201). Cresskill, NJ: Hampton Press.

Georgesen, J. C., & Harris, M. J. (1998). Why's my boss always holding me down? A meta-analysis of power effects on performance evaluations. *Personality and Social Psychology Review*, *2*, 184–195.

Gold, C., Voracek, M., & Wigram, T. (2004). Effects of music therapy for children and adolescents with psychopathology: A meta-analysis. *Journal of Child Psychology and Psychiatry*, *45*, 1054–1063.

Hall, J. A., Halberstadt, A. G., & O'Brien, C. E. (1997). "Subordination" and nonverbal sensitivity: A study and synthesis of findings based on trait measures. *Sex Roles*, *37*, 295–317.

Hamilton, M. A., & Hunter, J. E. (1998). The effect of language intensity on receiver evaluations of message, source, and topic. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 99–138). Cresskill, NJ: Hampton Press.

Hamilton, M. A., & Mineo, P. J. (2002). Argumentativeness and its effect on verbal aggressiveness: A meta-analytic review. In M. Allen, R. W. Preiss, B. M. Gayle, & N. A. Burrell (Eds.), *Interpersonal communication research: Advances through meta-analysis* (pp. 281–314). Mahwah, NJ: Lawrence Erlbaum Associates.

Hogben, M. (1998). Factors moderating the effect of televised aggression on viewer behavior. *Communication Research*, *25*, 220–247.

Irvin, J. E., Bowers, C. A., Dunn, M. E., & Wang, M. C. (1999). Efficacy of relapse prevention: A meta-analytic review. *Journal of Counseling and Clinical Psychology*, *67*, 563–570.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*, 681–706.

Khaleque, A., & Rohner, R. P. (2002). Perceived parental acceptance–rejection and psychological adjustment: A meta-analysis of cross-cultural and intracultural studies. *Journal of Marriage and Family*, *64*, 54–64.

Kierein, N. M., & Gold, M. A. (2000). Pygmalion in work organizations: A meta-analysis. *Journal of Organizational Behavior*, *21*, 913–928.

Kim, M. S., & Hunter, J. E. (1993a). Attitude–behavior relations: A meta-analysis of attitudinal relevance and topic. *Journal of Communication*, *43*, 101–142.

Kim, M. S., & Hunter, J. E. (1993b). Relationships among attitudes, behavioral intentions, and behavior: A meta-analysis of past research, part 2. *Communication Research*, *20*, 331–364.

Knight, G. K., Guthrie, I. K., Page, M. C., & Fabes, R. A (2002). Emotional arousal and gender differences in aggression: A meta-analysis. *Aggressive Behavior*, *28*, 366–393.

Kossek, E. E., & Ozeki, C. (1998). Work–family conflict, policies, and the job–life satisfaction relationship: A review and directions for organizational behavior-human resources research. *Journal of Applied Psychology*, *83*, 139–149.

La France, B. H., Heisel, A. D., & Beatty, M. J. (2004). Is there empirical evidence for a nonverbal profile of extraversion?: A meta-analysis and critique of the literature. *Communication Monographs*, *71*, 28–48.

Le, B., & Agnew, C. R. (2003). Commitment and its theorized determinants: A meta-analysis of the investment model. *Personal Relationships*, *10*, 37–57.

Levine, T. R., Bresnahan, M. J., Park, H. S., Lapinski, M. K., Wittenbaum, G. M., & Shearman, S. M. et al. (2003). Self-construal scales lack validity. *Human Communication Research*, *29*, 210–252.

Marcus-Newhall, A., Pedersen, W. C., Carlson, M., & Miller, N. (2000). Displaced aggression is alive and well: A meta-analytic review. *Journal of Personality and Social Psychology*, *78*, 670–689.

Murnen, S. K., & Stockton, M. (1997). Gender and self-reported sexual arousal in response to sexual stimuli: A meta-analytic review. *Sex Roles*, *37*, 135–153.

Murnen, S. K., Wright, C., & Kaluzny, G. (2002). If "boys will be boys," then girls will be victims? A meta-analytic review of the research that relates masculine ideology to sexual aggression. *Sex Roles*, *46*, 359–375.

Preiss, R. W., & Allen, M. (1998). Performing counterattitudinal advocacy: The persuasive impact of incentives. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 241–242). Cresskill, NJ: Hampton Press.

Rafaeli-Mor, E., & Steinberg, J. (2002). Self-complexity and well-being: A review and research synthesis. *Personality and Social Psychology Review*, *6*, 31–58.

Riggio, H. R., & Riggio, R. E. (2002). Emotional expressiveness, extraversion, and neuroticism: A meta-analysis. *Journal of Nonverbal Behavior*, *26*, 195–218.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin*, *114*, 363–375.

Schutz, H., & Six, B. (1996). How strong is the relationship between prejudice and discrimination?: A meta-analytic answer. *International Journal of Intercultural Relations*, *20*, 441–462.

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, *15*, 325–343.

Sherry, J. L. (2001). The effects of violent video games on aggression: A meta-analysis. *Human Communication Research*, *27*, 409–431.

Silverman, I. W. (2003). Gender differences in delay of gratification: A meta-analysis. *Sex Roles*, *49*, 451–463.

Stith, S. M., Rosen, K. H., Middleton, K. A., Busch, A. L., Lundeberg, K., & Carlton, R. P. (2000). The intergenerational transmission of spouse abuse: A meta-analysis. *Journal of Marriage and the Family*, *62*, 640–654.

Tamres, L. K., Janicki, D., & Helgeson, V. S. (2002). Sex differences in coping behavior: A meta-analytic review and an examination of relative coping. *Personality and Social Psychology Review*, *6*, 2–30.

Timmerman, L. M. (2002). Comparing the production of power in language on the basis of sex. In M. Allen, R. W. Preiss, B. M. Gayle, & N. A. Burrell (Eds.), *Interpersonal communication research: Advances through meta-analysis* (pp. 73–88). Mahwah, NJ: Lawrence Erlbaum Associates.

Whitely, B. E., Jr., Nelson, A. B., & Jones, C. J. (1999). Gender differences in cheating attitudes and classroom cheating behavior: A meta-analysis. *Sex Roles*, *41*, 657–680.

Wolfe, D. A., Crooks, C. V., Lee, V., McIntyre-Smith, A., & Jaffe, P. G. (2003). The effects of children's exposure to domestic violence: A meta-analysis and critique. *Clinical Child and Family Psychology Review*, *6*, 171–187.