FEATURE ARTICLES

Outcome-Reporting Bias in Education Research

Therese D. Pigott[1], Jeffrey C. Valentine[2], Joshua R. Polanin[3], Ryan T. Williams[4], and Dericka D. Canada[5]

Outcome-reporting bias occurs when primary studies do not include information about all outcomes measured in a study. When studies omit findings on important measures, efforts to synthesize the research using systematic review techniques will be biased and interpretations of individual studies will be incomplete. Outcome-reporting bias has been well documented in medicine and has been shown to lead to inaccurate assessments of the effects of medical treatments and, in some cases, to omission of reports of harms. This study examines outcome-reporting bias in educational research by comparing the reports of educational interventions from dissertations to their published versions. We find that nonsignificant outcomes were 30% more likely to be omitted from a published study than statistically significant ones.

Keywords: communication; experimental research; meta-analysis; program evaluation; research methodology; research utilization

Outcome-Reporting Bias

A number of factors influence the validity of a meta-analysis. Among these are the accuracy and completeness of both the literature search and the studies contributing to the analysis. One well-documented threat to the validity of meta-analytic results is publication bias, which refers to the tendency for studies lacking statistically significant effects to go unpublished (see Rothstein, Sutton, & Borenstein, 2005, for a review). When the primary outcome in a study is not statistically significant, study authors are less likely to submit the paper for publication at all (Cooper, DeNeve, & Charlton, 1997) and if they do submit it, take longer to do so (Suñé, Suñé, & Montoro, 2013). And even if submitted, papers lacking statistically significant primary outcomes are less likely to be published (Hopewell, Loudon, Clarke, Oxman, & Dickersin, 2009). Moreover, all else being equal, studies with "less" statistical significance have smaller effect sizes. Given that these studies are less likely to be published, a review might present an overly optimistic (i.e., positively biased) picture of the evidence if the review relies on mostly published sources.

Of course, the degree to which the primary studies are accurately and completely reported is also relevant to the validity of a review. If primary researchers do not provide a full and accurate reporting of the study's methods and results, then inferences from the review will likely also be biased (Orwin & Cordray, 1985). One specific problem of incomplete reporting focuses on outcomes measured in a study. Outcome-reporting bias refers to omitting from primary study reports outcomes that were actually collected. Thus, outcome-reporting bias can result when primary researchers either incompletely report outcomes gathered or omit entirely any mention of particular outcomes and can be difficult to detect once a study has been written up in final form. At best, a study might report that an outcome was measured and then report that the outcome was not statistically or substantively significant. Alternatively, researchers may choose not to report an outcome and omit all mention of its measurement. If outcomes have been censored from study reports because of their results (e.g., the findings are not statistically significant), then conclusions drawn from the incomplete evidence are potentially biased.

Outcome-Reporting Bias in Medical Research

The existence of outcome-reporting bias has been documented in medical research by comparing the protocols for a trial with the published results. Protocols, developed before conducting the study, provide operational details on the study's methods and analysis plan. Chan, Hróbjartsson, Haahr, Gøtzsche, and Altman (2004) collected the protocols of randomized trials reviewed by two scientific-ethical committees (similar to institutional review boards) in Denmark. The researchers compared the outcomes

[1]Loyola University Chicago, IL
[2]University of Louisville, KY
[3]Vanderbilt University, Nashville, TN
[4]The University of Memphis, TN
[5]Boston College, MA

Educational Researcher, Vol. 42 No. 8, pp. 424–432
DOI: 10.3102/0013189X13507104

reported in the protocols with the outcomes reported in published reports and found evidence of outcome-reporting bias. For example, 71% of statistically significant outcomes were reported versus 56% of nonsignificant findings, resulting in an odds ratio of 2.4 (i.e., the odds of an outcome being reported were 2.4 times greater for statistically significant outcomes than the odds for nonstatistically significant outcomes). Chan, Krleza-Jeric, Schmid, and Altman (2004) found similar results in comparing the protocols for trials funded by the Canadian Institutes of Health Research with their published reports.

Turner, Matthews, Linardatos, Tell, and Rosenthal (2008) extended the research base on outcome-reporting bias by examining the Federal Drug Administration's reviews of 12 antidepressant agents along with the matching published reports on these drugs. Overall, the researchers found a bias toward the publication of positive results. Most troubling, a meta-analysis using only the published reports found an average effect size that was larger than a meta-analysis using the unpublished FDA reviews. Similarly, Vedula, Bero, Scherer, and Dickersin (2009) compared unpublished reports of trials of off-label indications of the drug gabapentin with published reports of these same trials. The unpublished reports were internal documents that were obtained during the course of a lawsuit against two pharmaceutical companies. The primary outcome was changed in many published reports with secondary outcomes becoming primary, or reports of the primary outcome omitted (a process also observed by Chan et al., 2004). To illustrate why this is a problem, assume that a researcher is studying the effects of a reading intervention and measures reading achievement (the primary outcome) and self-esteem (a secondary outcome). On finding that the results are not statistically significant for reading achievement but are statistically significant for self-esteem, the researcher frames the intervention for publication as one that improves self-esteem and cites literature linking self-esteem to reading achievement. Changing the stated nature of a study's intent because of the results does not provide an accurate representation of the research.

Furthermore, Vedula et al. (2009) found that the published reports tended to indicate fewer adverse indications for gabapentin than the unpublished internal documents. To extend the reading intervention analogy, imagine that the researcher found that students receiving the intervention had lower reading motivation, but then failed to report this result. Clearly, such omissions have troubling implications for understanding what the totality of the evidence says about the effects of an intervention.

Outcome-Reporting Bias in Education and the Social Sciences

Few studies have been conducted in the social sciences to investigate the prevalence of outcome-reporting bias. One recent study focused on the behavior of researchers with regard to questionable reporting practices. John, Loewenstein, and Prelec (2012) anonymously surveyed more than 2,000 psychologists working at research universities in the United States and found that 63% admitted to not reporting all dependent measures that they assessed. Estimates were even higher in a condition that was incentivized to tell the truth. This finding is consistent with Chan and Altman's (2005) research that included a retrospective analysis of published randomized trials and a follow-up survey of the trials' authors. Combining data from the author surveys and the publications, Chan and Altman found that 75% (380/505) of trials did not fully report all their efficacy outcomes in the journal publication. Of the 308 trials that measured potential harms of a treatment, 64% did not fully report their harm outcomes. When asked for reasons why outcomes were not reported, 24% cited lack of statistical significance, with journal space restrictions given as the reason by 47% of the authors. For harm outcomes, 50% of the authors cited lack of statistical significance.

One other relevant line of research in the social sciences is Orwin and Cordray's (1985) study of deficient reporting in primary studies. These authors focused on the impact of incomplete reporting on subsequent meta-analysis results. They hypothesized that deficient reporting practices in a primary study would lead to greater uncertainty by coders of a research review and would ultimately affect the conclusions drawn from the review. Orwin and Cordray assessed the relationship between interrater reliability of codes in a meta-analysis and the confidence that coders felt in assessing the studies. Confidence ratings were positively related to the reliability of codes used in a meta-analysis. When coders were less confident about the information given in a report, their coding performance was less reliable, adding extraneous error into the meta-analysis. Thus, studies that either omit outcomes entirely or do not fully report on important outcomes can lead to biased and incomplete inferences.

As seen in the research described above, outcome-reporting bias exists as documented by direct evidence from comparing protocols to published research, and by indirect evidence from surveys of both medical and psychological researchers. Although some outcome-reporting bias may be driven by space limitations in journals, both Chan and Altman (2005) and John et al. (2012) both provide evidence that lack of statistical significance may be an important reason why outcomes are omitted in published reports. This latter finding highlights cynical interpretations of the processes underlying selective outcome reporting. Specifically, authors often benefit from publishing a study. Sometimes publication carries with it the possibility of direct financial reward (e.g., an intervention that, if deemed efficacious, can be licensed and sold), or the promise of more success in obtaining external funding in the future. Most incentives, like a perceived increase in the probability of obtaining tenure, are not so direct and strong, but this is not to say that they do not have the potential to affect behavior. Especially if authors believe that journal editors and reviewers have a preference for statistically significant findings (Cooper et al., 1997), then authors may have an interest in omitting non–statistically significant outcomes.

We contended that reviews based on studies that included censored outcomes were "potentially biased" because it is not immediately clear whether selective outcome reporting will result in a bias. There are at least three reasons why it might not bias the results (at least on average). First, although we think it unlikely, it is possible that selective outcome reporting is functionally a random process. Chan and Altman's (2005) research suggested that this was not a very tenable assertion, but there

may not be sufficient evidence to rule out this explanation completely. Second, if censored outcomes are strongly correlated with reported outcomes, then any resulting bias is likely to be minimal. For example, if a study collects two measures of self-esteem and reports the one that reveals a statistically significant relationship, it is likely that the difference in effects between the reported and the unreported self-esteem measure are small. Third, if the only outcomes censored are considered of less clinical or substantive importance, then it is unlikely that outcome censoring will have a significant impact on the results of a review. However, note that the Chan and Altman study found evidence of selective outcome reporting among outcomes labeled "important"—they just found more of it for outcomes that were perceived to be less important.

Given the documented presence of outcome-reporting bias in medical and social science research, we are interested in understanding the extent to which this phenomenon occurs in educational research. Investigating the potentially biasing effects of selective outcome reporting is, unfortunately, a much harder task in education and the social sciences than it is in medicine, where published research protocols are common. When protocols are available, researchers studying outcome-reporting bias can directly compare the study's initial protocol with the published versions of the trials. Unfortunately for education researchers (and their consumers), there is no analogous system for most studies. Institutional human subjects review boards (IRBs) do keep records of social science research, and could potentially be adapted to serve this and similar needs. Cooper et al. (1997), for example, examined publication bias by surveying researchers with an approved IRB protocol to see what happened to the research (e.g., whether it was abandoned, published, or completed but not published). However, IRB protocols are prepared essentially for nonexperts and therefore usually do not lay out in operational detail many of the methodological and statistical choices facing researchers, or even list all of the outcomes and how they will be measured. Given that each institution's IRB is also subject to localized practices and policies, the nature and completeness of these protocols would likely differ too much across institutions to provide any meaningful comparisons from protocol to published work.

Education researchers, however, use two systems that could serve as the basis for an analysis of outcome-reporting bias. First, grant proposals play a role similar to that of a research protocol in that they describe (in generally highly operational terms) the methods and analytic strategies that will be employed. Unfortunately, collecting these would likely be time consuming and require consent from researchers, funding agencies, and/or others. For example, Spybrook and Raudenbush (2009) examined the precision and technical accuracy of the first wave of randomized trials funded by the Institute of Education Sciences. To obtain the proposals funded by IES, they wrote directly to the investigators and received 40 of 55 proposals. For the remaining 15 proposals, they filed a Freedom of Information Act request to IES in 2006. At the time of publication in 2009, they still had not received any of the 15 requested proposals.

The second system that could be used is the dissertation process, which has notable advantages for researchers interested in examining outcome-reporting bias. Dissertations are typically not approved on the basis of their results (i.e., a failure to achieve statistical significance usually has no impact on a committee's decision regarding whether or not to approve the work). Most universities, furthermore, follow a process in which dissertation research is first proposed (and approved by a committee) then carried out. Normative understandings of the dissertation process (e.g., that it is developmental) and lack of length limitations generally lead to works that are reported in more detail than are typical journal articles. Though dissertation proposals are not available in the public domain, the dissertation itself usually presents a complete record of the methods and procedures that were actually used in the study, given the many changes that can occur between the proposed research and its actual implementation. Although outcome-reporting bias may also occur between the proposed dissertation research and the final dissertation, students are less likely to be constrained by the reasons cited by the authors of published papers such as space limitations or a bias toward statistically significant results. Finally, a large percentage of dissertations are available via electronic databases (e.g., ProQuest Digital Dissertations) to which many university libraries subscribe, and as such are easily retrievable.

The goals of this study, therefore, were to examine whether outcome-reporting bias exists in educational research, to estimate the magnitude of that bias, and to explore whether this effect is moderated by identifiable contextual variables. We choose education research as a field as we were interested in outcome-reporting bias in intervention research, a field both similar to the medical studies that have been conducted on outcome-reporting bias, and of particular interest to our work on the meta-analysis of intervention studies in education. We located dissertations conducted in education (broadly considered), searched for published versions of these dissertations, and compared the measured outcomes in the dissertation to the measured outcomes reported in the published version. Below we outline in more detail the research methods used, our results, and provide some suggestions for future research and for reporting standards in the social sciences.

Methods

Research Universities

To obtain our sample of dissertations, we focused on the 96 research universities designated by the Carnegie classification as very high research activity (RU/VH) universities as of 2005 (McCormick & Zhao, 2005). We were interested in the subsequent published versions of a dissertation, and we assumed that graduates of RU/VH institutions were more likely to pursue academic careers than students enrolled in other types of doctoral institutions and hence have greater motivation for pursuing publication (e.g., more than half of all Education doctorates in a given year are awarded by RU/VH institutions; National Science Foundation, 2011). All of the 96 institutions designated as RU/VH in 2005 were included in this study.

Search Strategy and Information Retrieval

The first stage of this study involved a comprehensive search for dissertations focused on education completed at the 96 RU/VH

universities. We searched the ProQuest Dissertations and Theses Database between the years 2001 and 2005 inclusive. There were two reasons for these specific date limits in the search. First, we were interested in inferences about the current state of the problem of data censoring and thus limited the lower date limit to 2001. Second, limiting the search to 2005 provided ample time for the dissertations to cycle through the entirety of the peer review and publication process. *Education* was used as a keyword and *Ph.D. or Ed.D.* was used to specify the degree earned. The process was repeated for each of the RU/VH universities. All results generated were saved for title and abstract screening.

To provide some focus and context for our findings, we limited our search to studies that investigated the effect of some educational intervention on student outcomes. As such, dissertations were retained for in-depth screening and analysis if the title and/or abstract indicated that the author provided an educational intervention for students in pre-kindergarten (pre-K) to Grade 12. We focused on randomized experimental studies and quasi-experimental studies where the goal was to arrive at an estimate of the treatment effect on a set of outcomes. Both multigroup and single-group experimental studies (pretest-posttest studies) were included. By narrowing our search to interventions, we excluded observational studies where the analysis may include more complex models and could lead to more difficulties in identifying the primary goal of the study. We imposed the limits of pre-kindergarten through Grade 12 because we assumed that these studies would be more conceptually similar (i.e., focused on an academic or behavioral intervention during the years most children must attend school) and likely to take place in an educational setting. All dissertations were screened by at least two individuals working independently. Disagreements or ambiguities that arose during the screening process were discussed until a consensus was achieved.

For each dissertation that met these initial inclusion criteria, we subsequently searched for published versions of the same study. Google Scholar was the primary search tool utilized for this process although both PsycINFO and ERIC were also used. For each dissertation, combinations of the title, keywords, and author name were used to locate the publication. In most cases, it was easy to match dissertations to publications based on the title and author. We assumed that the student would be an author on the published version of the paper and also checked that the samples appeared to be the same. In cases where we were unsure of a match, we used the full text of the dissertation and article to reach a conclusion. We collected all references to the dissertation, and none of the dissertations were represented by more than one subsequently published study. Thus, our analysis focused on the dissertation–publication pair. The search of the published version of the dissertation concluded in May 2011.

Because we were interested in the extent to which statistical significance might be related to outcome-reporting bias, our final inclusion criteria for dissertations and their subsequent published version relate to hypothesis testing. We focused on substantive (e.g. educational, social, and psychological) outcomes that were hypothesized to be affected by the intervention. Ancillary hypothesis tests such as baseline equivalence, normality, correlational, and homogeneity tests were excluded because it was not clear to us that outcome-reporting bias would work in the same way for these types (e.g., for tests of baseline equivalence, authors might be motivated to censor statistically significant results). The outcomes must have been formally tested in a manner that allowed the extraction or calculation of a p-value because we were interested in seeing how statistical significance relates to the probability of publication. Thus, dissertations using single-case experimental designs were not considered in this study. Furthermore, the analytic procedures and outcome measures must have remained constant in both papers. For example, if a dissertation conducted a series of t-tests and the publication used multivariate analysis of covariance (MANCOVA) on the same outcomes, it would have been excluded because we were interested in how the original analysis was reported in the published version.

Coding

Study pairs that met all inclusion criteria were coded at the study level and outcome level. Study-level codes included dissertation and publication year, publication source, and sample size, and type of intervention used. Outcome-level codes included p values and statistical test used. Outcomes coded for analysis were those that specifically reflected the intervention outcomes as they pertained to the pre-k through Grade 12 sample participants. This included all main and interaction effects of the intervention as well as any subgroup analyses that were reported in the dissertation. Table 1 provides an example of how we gathered the information about statistical tests within each dissertation and matching published paper.

Analysis

In order to compute the overall odds ratio across all matched dissertation and published paper pairs, we used the Mantel–Haenszel meta-analytic approach to estimate a weighted average odds ratio and a weighted average risk ratio (Shadish & Haddock, 2009). The studies included in this analysis varied considerably in the number of significance tests they conducted and we wanted to incorporate this variation in our estimation procedures. We estimated a weighted average effect size using the Mantel–Haenszel method, stratifying by study. For $i = 1, \ldots k$ dissertation–paper pairs, the Mantel–Haenszel mean odds ratio (Shadish & Haddock, 2009) is given by

$$\overline{OR}_{MH} = \frac{\sum_i^k a_i d_i / n_i}{\sum_i^k b_i c_i / n_i}, \quad (1)$$

where the cell counts for each pair i are defined in Table 1. We used the Robins–Breslow–Greenland variance estimator (Robins, Breslow, & Greenland, 1986) for the estimated Mantel–Haenszel odds ratio, and the Greenland–Robins variance estimator for the estimated Mantel–Haenszel risk ratio (Greenland & Robins, 1985). Alternatively, we attempted to compute study-level odds ratios in a meta-analysis. This method was compromised by a large number of zero cell counts, indicating that many outcomes did not appear in the subsequent published report. The

Table 1
Sample Data Collected for Each Dissertation and Matched Published Paper

	Reported in published study		
Reported in dissertation	Reported	Not reported	Totals
Statistically significant	a_i	b_i	$a_i + b_i$
Not statistically significant	c_i	d_i	$c_i + d_i$
Totals	$a_i + c_i$	$b_i + d_i$	n_i

Note. Cells are identified by letters. The number of statistically significant outcomes reported in a dissertation and also reported in the published version are reported in cell *a*. See equation (1).

Table 2
Descriptive Statistics for Tests in Dissertation and Published Paper

Publication version	Minimum number of tests	Maximum number of tests	Median	IQR	*M*	*SD*
Dissertation	2	173	14	18	20.24	23.69
Published	0	96	7	10	10.16	13.60

Note. $N = 79$ dissertation–published paper matched pairs. IQR = interquartile range.

Mantel–Haenszel method obviates this issue and provides a defined weighted average so long as each cell is not uniformly zero across each of the samples.

Results

From the 96 institutions, we identified 9,530 dissertations. Of these, we found 621 dissertations (6.5%) that reported on an educational intervention with prekindergarten through 12th-grade students. Of the 621 dissertations on education interventions, we identified 79 that were subsequently published (12.7% of the dissertations on interventions).

Within the 79 studies, we found 1,599 different treatment outcomes. Table 2 provides the descriptive statistics for the numbers of statistical tests identified in the dissertations and subsequent published papers. Overall, 46% of the statistical tests reported in the dissertations were statistically significant, underscoring a serious deficiency in statistical power across this sample of studies. On average, the published version of the dissertation included about half of the outcomes reported in the dissertation. Of the 79 published reports, 19 (24%) included all of the outcomes described in the dissertation.

Next we examined, among all of the individual outcomes collected, the (unweighted) estimated probability of being published and not being published for significant and nonsignificant outcomes. For outcomes reported in the published version of the studies, 54% were statistically significant and 46% were not statistically significant. Looking at these probabilities alone, outcome reporting in education would not look so problematic. However, examining published outcomes alone ignores the information that we were able to glean from the dissertation versions of the studies. If we examine unpublished outcomes only (i.e., outcomes measured in the dissertations but not reported in the published versions), 35% were statistically significant and 65% were not statistically significant. The discrepancy in the proportion of statistically significant results across the outcomes that were published and those that went unpublished is our first indication that there might be systemic outcome-reporting bias in these studies. Overall, 36 of 79 dissertations (46%) appeared to experience some outcome censoring because of statistical significance.

We next computed the Mantel–Haenszel odds ratio, which allowed us to weight studies by the number of statistical tests that were reported in the dissertation. The mean odds ratio was 2.41 (with a 95% confidence interval ranging from a low of 1.79 to a high of 3.25). Thus, the odds of a statistically significant outcome in a dissertation appearing in the published version were 2.41 greater than the odds of a nonstatistically significant outcome appearing in the published version.

There are several equally accurate ways of framing this effect. One is to present these results as a risk ratio, which in this case describes the risk of an outcome being omitted from the published report conditional on its statistical significance. Here, the risk ratio was 1.30, meaning that nonsignificant outcomes were 1.30 times (or 30%) less likely to appear in the published version than were statistically significant outcomes. Conversely, changing the focus to nonpublication, the risk ratio becomes .78, indicating that significant dissertation outcomes are about 22% less likely to be omitted from publication compared to nonsignificant outcomes. Another way of framing this effect is to say that the probability of an outcome being published, conditional on statistical significance in the dissertation, was .71, whereas the probability of being published, conditional on nonsignificance in the dissertation, was .29.

We could not reject the null hypothesis of homogeneity for this set of effect sizes, $Q(78) = 72.3$, $p = .66$. Regardless, we were still interested in whether we could identify any contextual variables that might moderate the overall weighted mean effect size.

Table 3
Subgroup Analyses

Variable	Subgroup	OR	95% CI (lower, upper)	$\chi^2(1)$, *p*
Time to publication				1.62, .20
	Within 2 years	2.00	1.34, 3.00	
	More than 2 years	2.97	1.89, 4.66	
Sample size				0.53, .47
	<100	2.35	1.65, 3.36	
	≥100	3.05	1.66, 5.55	
Number of statistical significance tests				0.51, .48
	≤14	2.68	1.76, 4.08	
	>14	2.16	1.41, 3.30	
Type of outcome measure				0.37, .55
	Academic achievement	2.46	1.71, 3.52	
	Socioemotional-behavioral	1.99	1.13, 3.53	

Note. OR = odds ratio; CI = confidence interval.

To do this, we conducted a series of subgroup analyses. We looked at four classes of subgroups: Time to publication (within 2 years vs. more than 2 years); sample size (less than 100 or 100 or more); the number of significance tests conducted in the dissertation (14 or fewer vs. more than 14); and the type of outcome analyzed (academic achievement vs. socioemotional-behavioral). The categories for the first three potential moderators were determined by a median split, as the distributions were decidedly nonnormal. Not one of these moderator analyses was statistically significant (all *p*'s > .20; see Table 3), meaning that the mean weighted odds ratio of 2.41 remains our best estimate of the extent of outcome-reporting bias in these studies.

Discussion

The goal of this project was to investigate the outcome-reporting tendencies of dissertation authors. We hypothesized that authors publish statistically significant outcomes more often than statistically nonsignificant findings. Our results support this hypothesis. Furthermore, the results of the subgroup analyses revealed that the overall weighted odds ratio was not conditional on time to publication, sample size, the number of significance tests conducted, or outcome type. We also found that, on average, 46% of the statistical tests of treatment outcomes reported in dissertations resulted in a rejection of the null hypothesis. This finding suggests that statistical power in dissertations in educational research is approximately in line with, if perhaps a bit lower than, other estimates of typical statistical power in the social and behavioral sciences (e.g., Bezeau & Graves, 2001; Cohen, 1962; Ioannidis 2005; Rossi, 1990; Sedlmeier & Gigerenzer, 1989).

Implications

We expected to find evidence of biased outcome reporting, and we were able to document its presence in this population of dissertations that were completed between 2001 and 2005 and subsequently appeared by 2011 in a published version. This result fits into an emerging research base raising concern about the degree of flexibility researchers have in designing and analyzing studies, the lack of transparency of the reporting of many of these choices, and the sometimes dramatic impact these choices can have on study results (e.g., Francis, in press; Ioannidis, 2005; John et al., 2012; Simmons, Nelson, & Simonsohn, 2011).

Given that the mean number of treatment outcome tests reported in dissertations exceeded 20, it is little wonder that some of these go unreported in the published articles. Clearly the processes underlying the selective reporting of outcomes needs further study. One reason for selectively omitting nonsignificant findings may be a lingering misinterpretation of *p* values (i.e., that if a result was not statistically significant it means that "nothing interesting" was found). Because most studies are not conducted with high degrees of statistical power, a researcher may not obtain statistical significance simply because the study was too small relative to the population effect being measured. Low statistical power is one reason why statistical significance and substantive importance should not be conflated (Valentine, Pigott, & Rothstein, 2010).

In addition, virtually all of the dissertation authors in our study appear to have gone on to academic careers. The pressures of this career path may lead authors (and perhaps their advisors) to engage in motivated reasoning regarding the reporting of the analyses. Researchers may also be motivated by a desire to present what appears to be a more coherent report of the study, or stated differently, authors may be motivated to tell a good story. Doing so may involve a process of "sharpening" the perceived important results (i.e., the statistically significant ones) and "leveling" the results perceived to be less important (i.e., the non–statistically significant ones). Finally, we did not code for whether the statistical tests were of main effects or of interactions, and it may be that selective reporting is more likely among interaction tests (e.g., dropping subgroup analyses that were not statistically significant).

No matter what the reason for incomplete reporting is, the result is a biased picture of the research findings. Even if, for example, our results were solely a function of authors not reporting nonsignificant subgroup analyses, it still implies that the reporting of subgroup analyses is conditional on statistical significance, and the result is an incomplete understanding of both the intended and unintended effects of a treatment. As in the studies on biased outcome reporting in drug trials, we may reach erroneous conclusions on the relative effectiveness or ineffectiveness of a treatment.

Finally, we would be remiss not to acknowledge that some skeptics may use our study as a reason for not trusting the results of research syntheses and meta-analyses. We think this is an overstatement. Our motivation for undertaking this study was to increase the value and accuracy of methods for synthesizing literature, but outcome-reporting bias has the potential to affect any review, regardless of the specific methods used to synthesize studies. It also affects the interpretation of any single study. The problems raised by outcome-reporting bias do not go away by choosing something other than meta-analysis as the synthesis technique, or by avoiding synthesis altogether. Encouraging researchers to report all findings can only contribute to the knowledge base on interventions and other phenomena in education.

Suggestions

Many of the reasons for not fully reporting all outcomes of an intervention study relate to the constraints around the amount of information that can fit into a single published manuscript (Orwin & Cordray, 1985). Fortunately, the growing popularity of web-based storage for supplementary research materials makes comprehensive and accurate reporting more of a possibility than ever before. The full information about an intervention, including the reports of all outcomes measured, could be made public for research reviewers, leading to greater validity of literature syntheses.

Going a step further, researchers should be encouraged to archive their data from educational interventions for fair public use after the passage of a reasonable amount of time. The National Institutes of Health and the National Science Foundation both have policies for sharing data collected with funding from these institutions. The Interuniversity Consortium for Political and Social Research at the University of Michigan warehouses a number of public-use databases and is beginning to archive data from individual researchers in smaller-scale studies. Journals could facilitate this process by providing space for data warehousing, and by requiring a well-documented database be submitted for online archiving prior to accepting a study for publication. Well-documented data repositories only increase the accuracy and completeness of resulting meta-analyses. Even beyond the advantage of having the data available so that others can attempt to reproduce the results, the use of data warehousing could have additional benefits. For example, it would likely lead to increased use of Individual Participant Data meta-analyses, where data from original studies when available are combined with study-level data in a single systematic research review (Cooper & Patall, 2009; Pigott, Williams, & Polanin, 2012; Valentine & Thompson, 2012).

What steps can we take to increase the quality of reporting of primary studies, and thus the completeness of syntheses using these studies? Ideally, all educational research would start with a highly operational, publicly available protocol that guides the research. This is a lofty goal, but perhaps we can start with the development of professional norms that hold researchers responsible for fully documenting the research methods and analytic choices, including reporting all outcomes they measure. Further, we can continue to study the prevalence and magnitude of outcome-reporting bias by obtaining protocols of studies either from funding agencies or from local human subjects review boards. In addition, more research could be done investigating the processes underlying selective outcome reporting.

Finally, the educational research community has been slower in adopting reporting guidelines for published work than other disciplines. Medical clinical trials follow the CONSORT guidelines for reporting the procedures and results of medical interventions (Schulz, Altman, & Moher, 2010). The American Psychological Association also has published guidelines for reporting of empirical studies in journal articles (JARS), and of meta-analyses (MARS) (APA Publication and Communication Board Working Group on Journal Article Reporting Standards, 2008). For example, the journal article reporting standards request that authors report details regarding sampling procedures, sample characteristics (including both major demographic characteristics and baseline topic specific characteristics), and, critically, a description of all primary and secondary outcomes measured. The education research community could go one step further by requiring study authors to explicitly state whether they have engaged in certain "grey area" practices, such as excluding or trimming observations, testing multiple models (e.g., several different sets of covariates), and so on. And if researchers have engaged in grey area practices, they should be required to state explicitly the timing of and rationale for these choices, and should provide sensitivity analyses that disclose what would have happened had different choices been made. With regard to outcome-reporting bias in particular, consistency in reporting of all outcomes measured in a study would improve not only our understanding of the full range of potential outcomes of an intervention, but would also allow for a more complete picture of the state of a research area in a research synthesis.

REFERENCES

APA Publication and Communication Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839–851. doi:10.1037/0003-066X.63.9.839

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical Experimental Neuropsychology, 23*, 399–406. doi:10.1076/jcen.23.3.399.1181

Chan, A. W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *British Medical Journal, 330*, 753–760. doi:10.1136/bmj.38356.424606.8F

Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting

of outcomes in randomized trials. *Journal of the American Medical Association, 291*, 2457–2465. doi:10.1001/jama.291.20.2457

Chan, A. W., Krleza-Jeric, K., Schmid, I., & Altman, D. G. (2004). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal, 171*, 735–740. doi:10.1503/cmaj.1041086

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

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. doi:10.1037//1082-989X.2.4.447

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*, 165–176. doi:10.1037/a0015565

Francis, D. P. (in press). How easily can omission of patients, or selection amongst poorly-reproducible measurements, create artificial correlations? Methods for detection and implications for observational research design in cardiology. *International Journal of Cardiology*.

Greenland, S., & Robins, J. M. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics, 41*, 55–68. doi:10.2307/2530643

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews (1)*, MR000006. doi:10.1002/14651858.MR000006.pub3.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, 696–701. doi:10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–32. doi:10.1177/0956797611430953

McCormick, A. C., & Zhao, C. M. (2005). Rethinking and reframing the Carnegie classification. *Change, 37*, 50–57. doi:10.3200/CHNG.37.5.51–57

National Science Foundation. (2011). Doctoral recipients from US Universities (NSF13-301). Arlington VA: NSF. http://www.nsf.gov/statistics/sed/2011/start.cfm

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 97*, 134–147. doi:10.1037//0033-2909.97.1.134

Pigott, T. D., Williams, R. T., & Polanin, J. R. (2012). Combining individual participant and aggregate data in a meta-analysis with correlational studies. *Research Synthesis Methods*. Advance online publication. doi:10.1002/jrsm.1051

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42*, 311–323. doi:10.2307/2531052

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

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

Schulz, K. F., Altman, D. G., & Moher, D., for the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *British Medical Journal, 340*, 332–337. doi:10.1136/bmj.c332

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 257–293). New York, NY: Russell Sage Foundation.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi:10.1177/0956797611417632

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis, 31*, 298–318. doi:10.3102/0162373709339524

Suñé, P., Suñé, J. M., & Montoro, J. B. (2013). Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. *PLoS One, 8*, 1–8. doi:10.1371/journal.pone.0054583

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252–260. doi:10.1056/NEJMsa065779

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*, 215–247. doi:10.3102/1076998609346961

Valentine, J. C., & Thompson, S. G. (2012). Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods*. Advance online publication. doi:10.1002/jrsm.1064

Vedula, S. S., Bero, L., Scherer, R. W., & Dickersin, K. (2009). Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *New England Journal of Medicine, 361*, 1963–1971. doi:10.1056/NEJMsa0906126

AUTHORS

THERESE D. PIGOTT, PhD, is an associate dean and professor of research methodology at the School of Education, Loyola University Chicago, 820 N. Michigan Avenue, Chicago, IL 60611; *tpigott@luc.edu*. Her research focuses on methods for research synthesis and meta-analysis.

JEFFREY C. VALENTINE, PhD, is an associate professor at the University of Louisville, 309 CEHD, University of Louisville, Louisville, KY 40292; *jeff.valentine@lousiville.edu*. Most of his research involves using, explaining, and seeking to improve meta-analytic techniques.

JOSHUA R. POLANIN, PhD, is a postdoctoral fellow at Vanderbilt University (Peabody Research Institute, 230 Appleton Place, Nashville, TN 37203); *joshua.r.polanin@vanderbilt.edu*. His research focuses on improving the methods of meta-analysis and he is the managing editor of Campbell Collaboration's method's group.

RYAN T. WILLIAMS, PhD, is an assistant professor at The University of Memphis, 100 Ball Hall, The University of Memphis, 38152; *ryan.williams@memphis.edu*. His research focuses on developing quantitative methods for research synthesis and meta-analysis, applied measurement, and issues related to causal inference, design, and analysis.

DERICKA D. CANADA, MEd, is a diversity fellow and doctoral student in counseling psychology at Boston College; *dericka.canada@bc.edu*. Her research focuses on examining racial and ethnic identity through psychosocial and sociocultural lenses, as well as exploring the intersections of gender and racial identity, and racialized body image among Black women.

Manuscript received November 13, 2012
Revisions received March 18, 2013, and August 7, 2013
Accepted August 19, 2013