Journal of Universal Computer Science, vol. 14, no. 4 (2008), 575-589
submitted: 2/4/07, accepted: 13/2/08, appeared: 28/2/08 © J.UCS

Publication Bias in the Computer Science Education Research Literature

Justus J. Randolph & Roman Bednarik
(Department of Computer Science and Statistics,
University of Joensuu, Finland
justus@randolph.name and bednarik@cs.joensuu.fi)

Abstract: Publication bias is the tendency for investigations with primarily nonstatistically significant findings to be withheld from the research record. Because publication bias has serious negative consequences for research and practice, we gathered information about the prevalence and predictors of publication bias in the computer science education literature. From an initial random sample of 352 recent computer science education articles, we reviewed the 38 empirical articles that used inferential statistical analyses. We found that (a) the proportion of articles reporting primarily statistically significant findings in computer science education was very similar to the proportion in medical research, (b) that an article's having a female first author was a strong predictor of an article's having primarily statistically significant results, and (c) that there was a tendency for authors to emphasize statistically significant findings and deemphasize nonstatistically significant findings. Neither whether an investigation was reported in a journal or conference proceeding nor whether the source of funding was disclosed were significant predictors of an article's having statistically positive results.

Key Words: publication bias, computer science education, statistical reporting, gender

Category: K.3, K.3.2

1 Introduction

Reporting bias is a broad term that refers to systematic bias that affects which, and how rapidly, academic reports get published. Varieties of reporting bias include publication bias (i.e., positive results bias), language bias, funding bias, outcome selection variable bias, database bias, coding bias, citation bias, regional bias, developed country bias, among others [Cochrane Bias Methods Group n.d.]. In this article, we are interested in the phenomenon of publication bias (sometimes called *positive results bias*), which we define as *the tendency for articles with statistically positive results to get submitted and published and for articles with statistically negative or neutral results to not get submitted or published.*[1] Publication bias has been shown to distort the research record [Riniolo 1997] because "statistically [positive] results may dominate the research record and

[1] By *statistically positive* we mean that an article has more statistically significant results than nonstatistically significant results, by *statistically negative* we mean that an article has more nonstatistically significant results than statistically significant results, and by *statistically neutral* we mean that an article has an equal number of statistically significant and nonstatistically significant results.

skew the results of systematic reviews and meta-analyses in favor of treatments with [statistically] positive results" [Lee et al. 2006, p. 621]. Although publication bias in computer science education research probably will never lead to a loss of life, as it has in the field of medical science [Grimm 2006, p. 835], it could have serious consequences nonetheless. It could, for example, lead generations of computer science educators to adopt pedagogical strategies or tools that are not as effective as the research literature claims them to be. Also, because computer science education is a relatively young and emerging discipline, the systematic exclusion of research findings could cause subsequent generations of researchers to waste resources by conducting research that has already been done but not reported. In addition, some claim that it is "scientifically and ethically unacceptable to invite people to participate in these studies and not publish the results" [Grimm 2006, p. 837].

Publication bias comes in two forms: editorial bias and authorial bias. Editorial publication bias refers to the tendency of editors or reviewers to reject manuscripts that have statistically negative results, all other things being equal. Similarly, authorial bias refers to the tendency for authors to not submit manuscripts that have statistically negative or statistically neutral results, all other things being equal. Speculations on the cause of authorial bias include "a perceived lack of interest, methodological limitations, or the assumption that editors and reviewers are less likely to publish them" [Lee et al. 2006, p. 621]. Recent studies have shown that authorial publication bias is by far the most common of the two [Lee et al. 2006, Olson et al. 2002].

As a solution to the problem of publication bias, some specialty journals; like *PLoS Clinical Trials, The Journal of Universal Computer Science–The Forum for Negative Results*, or *The Journal for Negative Results in Biomedicine*; have adopted a policy that all high-quality articles will be accepted regardless of whether they report statistically positive results or not. However, many claim that such policies will lead to a proliferation of "so-what" studies and, as a result, journals that adopt those policies will not survive economically [Grimm 2006]. Others claim that the change must come from scientists themselves. We believe that a prerequisite for change is to make scientists aware of the issue of publication bias. In this study we investigate to what degree publication bias is present in our field of computer science education. With hope, those results will help spur awareness, increase knowledge, and eventually lead to behavior change in terms of reducing editorial and authorial bias.

In this investigation we were not able to isolate the unique effects of editorial publication bias because of the difficulty of identifying and gaining access to the thousands of rejected computer science education papers submitted between 2000 and 2005.[2] However, by looking at the proportions of published articles

[2] Some of the previous medical studies examined both accepted and rejected papers

that have statistically positive and statistically negative results, we can get a general picture of what publication bias, both editorial and authorial combined, looks like in the computer science education research literature. And, we can compare that picture to what it looks like in other fields, namely medicine, in which there is already much information about publication bias. In order to make comparisons between fields and to ground our methods, we first present the results of previous research on publication bias.

2 Previous Research

A search on February 2nd, 2007, of the ACM digital library using the the search phrase **publication bias** resulted in three results, none of which actually addressed the issue of publication bias. Also, we could not find a reference to *publication bias* in the index of Fincher and Petre's (2004) seminal book on computer science education research. Therefore, we are convinced that there is a lack of information about publication bias in the field of computer science education–and computer science in general[3].

Because of the absence of previous research on publication bias in the computer science education literature, we decided to look to the field of medicine. We looked to the field of medicine for two reasons. First, medical research has historically been a standard of comparison for education research [Riehl 2006]. (We consider computer science education research to be, mostly, a subset of education research.) Second, according to Riehl, "medical research parallels the scope of research pertinent to education, from neurological studies of brain functioning through sociological analysis of educational systems" (p. 24). Third, there has been a significant amount of high quality research on publication bias in the field of medicine and, therefore, we can use the results of those studies as a reference point for our own results.

To arrive at an estimate of the overall proportion of published, statistically positive studies in the medial research, we synthesized the results of a purposive sample of five major medical-research publication bias studies. Those five major studies were [Lee et al. 2006, Olson et al. 2002, Stern and Simes 1997, Easterbrook and Berlin 1991, Dickersin and Min 1993]. We chose Lee et al. and Olson et al. because of their currency and popularity–they were recently published and were discussed in detail in a influential article in *Science* [Grimm 2006].

so that the unique influence of editorial publication bias could be examined. Both Lee et al. [2006] and Olson et al. [2002] found no compelling evidence for editorial publication bias in the medical research record. Theoretically, a research record without editorial publication bias would have equal proportions of statistically positive rejected and accepted papers.

[3] However, a new forum that addresses the problem of publication bias in computer science is currently being established; it is called the *Forum for Negative Results*, a section in the *Journal of Universal Computer Science*. See [Prechelt 1997].

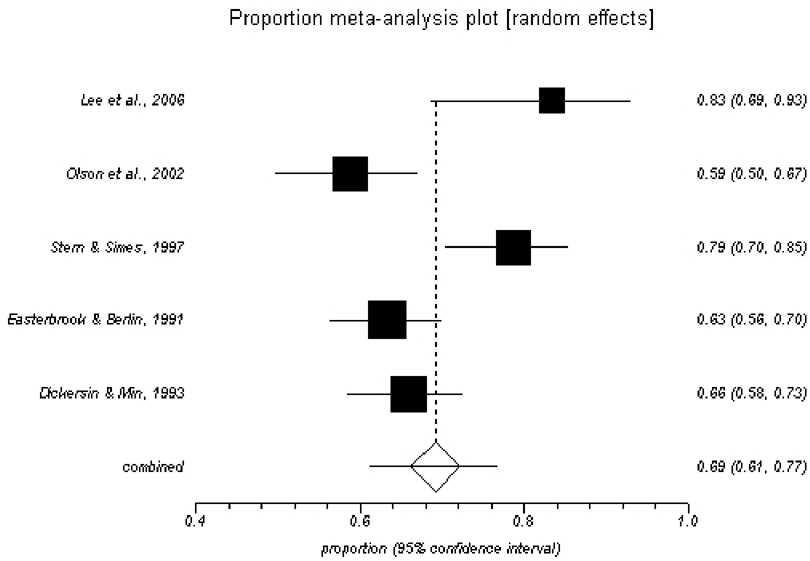

Figure 1: Forest plot of proportions of published, statistically positive articles in the medical research literature.

We chose Stern and Simes, Easterbrook and Berlin, and Dickersin and Min because of the large sample sizes of the published articles that they reviewed (n = 126, 207, and 184, respectively). We admit that this sample is far from comprehensive, but since the focus of this article is on computer science education research rather than medical research, we believe that this sample is adequate as a general point of reference for our own results.

Figure 1 shows the combined proportion for our primary outcome of interest– the percentage of published articles with statistically positive results. On average, 69.1% of the medical research studies reviewed had statistically positive results. The 95% confidence intervals around that proportion were 61.0% and 76.7%.

3 Research Questions

Given the lack of information about the prevalence of authorial and editorial publication bias in the computer science education literature and because of the harmful consequences of bias, we investigated the following research questions:

1. What are the proportions of statistically positive results, statistically negative results, and statistically neutral results in the articles published in the computer science education literature?

2. How do the proportions of statistically positive and statistically negative results compare to the proportions from medical research?

3. Do the proportions of statistically positive and statistically negatives results covary by type of publication (i.e., journal or conference proceedings), sex of first author, or whether funding is disclosed or not?

Investigating the proportions in our own field and in the field of medical research allows us to make comparisons between fields. If the proportions are similar, researchers will have some justification for generalizing the findings from the medical publication bias research to the computer science education research.

Because conference proceedings are regarded by many as having less merit than archival publications [National Research Council 1994], we investigated, whether computer science education research authors would tend to submit more manuscripts with statistically negative findings to conference proceedings than archival journals. (Randolph et al. [2007], however, found that despite commonly held perceptions, papers published in computer science education journals and conference proceedings had about the same level of methodological quality.) In case there was a tendency for funded authors to please funders by making statistically positive results public and keeping statistically negative results private, we investigated whether disclosure of funding source was a predictor of an article's having statistically positive results. Finally, because of the well-documented gender gap in computer science education, both at the student-level and at the instructor/researcher-level [Margolis and Fisher 2001, Selby et al. 1998], we investigated whether sex of first author was a predictor of an article's having statistically positive results.

4 Method

In this study, we replicated the methods of Olson et al. [Olson et al. 2002]. One difference between this study's method and Olson et al.'s method, however, was that we did not have access to articles that were not accepted for publication; we only examined articles that had been accepted. Our results therefore confound editorial and authorial bias. Nonetheless, we were still able to determine the prevalence of statistically positive articles, compare the prevalence between fields, and identify the predictors of an article's being statistically positive.

The sample on which our results were based was originally collected for [Randolph 2007]. Therefore, detailed information on the sampling procedure, criteria for inclusion and exclusion, sample characteristics, and a list of articles included in the total sample can be found there.

4.1 Sample

We took a random sample, stratified by year and forum, of 352 of the 1306 (26.9%) full papers published between 2000 and 2005 in eight of the major forums for publishing computer science education research. The eight major forums that were sampled from are listed below:

- *SIGCSE Bulletin,*
- *Computer Science Education, (a journal),*
- *The Journal of Computer Science Education Online,*
- *Proceedings of the Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education,*
- *Proceedings of the SIGCSE Technical Symposium on Computer Science Education,*
- *Proceedings of the Innovation and Technology in Computer Science Education Conference,*
- *Proceedings of the Australasian Computing Education Conference, and*
- *Proceedings of the International Computer Science Education Research Workshop.*

We excluded articles less than three pages long and nonpeer-reviewed articles. Of the 352 sampled articles, we only included the 40 articles that used inferential statistical analyses for primary outcomes. The other 312 articles in the sample consisted of 119 articles not dealing with human participants, 89 articles presenting only anecdotal evidence, and 104 articles that were empirical but did not use inferential statistical analyses for primary outcomes. Those 312 articles were not included in this analysis because the phenomenon of interest concerns bias towards statistically positive results. We did not consider a primary outcome to be whether control and treatment groups differed on a pretest, when a pretest-posttest with control group design was used. We excluded one conference paper that used inferential analyses on a primary outcome because it was a summary of a more comprehensive journal article that had already been included in the sample. We also excluded another paper because the reporting was insufficient to determine how many inferential tests had been conducted. In total then, 38 articles were included in the sample. Because we sampled these articles randomly, we assume that these 38 articles are representative of the population of recent computer science education articles that used statistical analyses. Although our sample is smaller than the samples in most of the medical articles, it is important to note that our sample accounts for about 27% of the estimated total population of empirical computer science education articles. Because our sampling ratio was 3.71 articles in the population to every 1 article sampled, we estimate that there was a population of 141 empirical computer science education articles, published between 2000 and 2005, in which statistical tests were conducted.

4.2 Coding scheme and procedure

For each of the 38 articles that used inferential analyses, we literally used the coding scheme reported in [Olson et al. 2002]:

We classified results as positive if they showed a statistically significant effect (p<.05; 95% confidence interval [CI] for difference excluding 0 or 95% CI for ratio excluding 1) on the primary outcome in that study and negative if they did not. If no primary outcome was stated or discernible, we classified results based on most outcomes. Results were unclear if they could not be classified as positive or negative, typical when many outcomes were reported and an equal number were positive and negative but none was primary. (p. 2826)

When an article reported multiple inference tests, we did not adjust the criterion of statistical significance (alpha); it was always the same regardless of the number of inference tests reported. In addition to coding the results of an article, we also coded whether the forum from which the article came was a journal or a conference proceeding, whether the first author was a female or a male, and whether a source of funding was disclosed. We considered *SIGCSE Bulletin*, *Computer Science Education*, and *The Journal of Computer Science Education Online* to be journals and the rest of the forums to be conference proceedings. We considered an article to have disclosed a funding source if there was an acknowledgment stating something to the effect that “this research was supported by a grant from. . . [some funding source].”

Two raters (the authors of this article) independently rated each of the 38 articles. Both raters were researchers at a department of computer science and statistics and had extensive previous experience with conducting methodological reviews of the computer science education literature. When raters disagreed on how an article should have been classified, they came to a consensus on which classification to use. Brennan and Prediger’s [1981] free-marginal kappa was used to estimate the level of agreement, before adjudication. Interrater reliabilities were not calculated for sex of first author, type of publication, or whether a source of funding was disclosed, because that was essentially factual information.

4.3 Data analysis

Binary logistic regression for predictors of statistically positive results was conducted using the method described in [Agresti 1996]. Logistic regression analyses were conducted using SPSS 11.0.1. We used StatsDirect software [StatsDirect 2005] to conduct meta-analyses of proportions. We used a random effects model for the meta-analyses summarized in Figure 1 and Figure 2, because Cochran's Q and the I^2 statistic (see [Higgins and Thompson 2002]) indicated that there was significant heterogeneity between studies. When doing regression analyses or meta-analyses, the statistically neutral category was excluded.

5 Results

5.1 Interrater reliabilities

There were initial disagreements on 11 of the 38 articles; the value of *kappa* was .57. Because there was such a large number of disagreements, we used an emergent coding technique to classify the reasons for the disagreements. The reasons that emerged were that (a) we disagreed on the total number of inferential tests that had been conducted, (b) we disagreed on which outcomes were primary, (c) we disagreed on the number of positive and negative tests because the positive results were expressed in tabular form or with mathematical symbols (e.g., "$p <$ .05") and negative results were expressed in written form without mathematical symbols (e.g., "None of the inferential tests were statistically significant"), and (d) we disagreed because of human error. One or more of the four reasons for disagreement could have been given as the cause of a disagreement. In whole or part, three disagreements were attributed to different inferences about the total number of tests, four disagreements were attributed to disagreement about primary outcomes, five disagreements were attributed to emphasized positive results and deemphasized negative results, and five disagreements were attributed to human error. When only taking into consideration the inferential tests that were reported in tabular form or set off with statistical symbols in the text (and therefore ignoring the inferential tests that were expressed without numerals), the percentage of articles with positive results was nearly 75%.

5.2 Main effects and predictors of positive results

Of the 38 articles, 22 were statistically positive (57.8%), 13 (34.2%) were statistically negative, and 3 (7.8%) were statistically neutral. Disregarding articles with statistically neutral results, 22 out of 35 articles (62.9%, with 95% confidence intervals of 45.9% and 78.1%) had statistically positive results. Since Cochran's Q and I^2 indicated heterogeneity between studies, we used random effects models for the meta-analyses summarized in Figure 1 and Figure 2.

Variable	B	SE B	df	Sig.	Exp(B)	Lower 95% C.I.	Upper 95% C.I.
Female first	2.485	1.13	1	0.03	12	1.32	108.78
Intercept	-0.87	0.42	1	0.84	0.92		

Table 1: Summary of Fitted Logistic Regression Equation for Predictors of Statistically Positive Results

Sex of first author	Stat. positive results	Stat. negative results	Total
Female	11	1	12
Male	11	12	23
Total	22	13	35

Table 2: Number of Articles with Statistically Positive Results Crosstabulated by Sex of First Author

The best-fitting logistic regression model [Agresti 1996] had only the intercept and sex of first author included. The omnibus test of model coefficients was statistically significant, $\chi^2(1, \text{N }(35) = 7.46, \text{p} = .006)$, which indicates that the chosen model was appropriate. That is, neither type of publication nor disclosure of funding were statistically significant predictors of an article's having statistically positive results.

Table 1 shows that the odds of a statistically positive article's having a female first author was 12 times greater than the odds of a statistically positive article's having a male first author. After discovering that sex of first author was the only statistically significant predictor of bias, we explored several possible confounding factors. The 12 female first authors were all different and all but two came from different institutions. Also, there was no obvious confound of region of first author's origin or forum where the article was published. Table 2 provides the numbers and percentage of articles that had statistically positive results when crosstabulated by sex of first author.

5.3 Comparison to medical research

Figure 2 shows again the forest plot of medical research studies on publication bias that was presented in Figure 1, but this time with the results of the current study also included. It shows that the proportion found in the current study falls in the lower end of the expected range of proportions of published, statistically positive articles found in medical publication bias articles. As expected from Figure 2, a logistic regression analysis showed that there was not a statistically significant difference between the current study and the group of medical studies, (odds ratio = 1.20, 95% CI, .59-2.43).

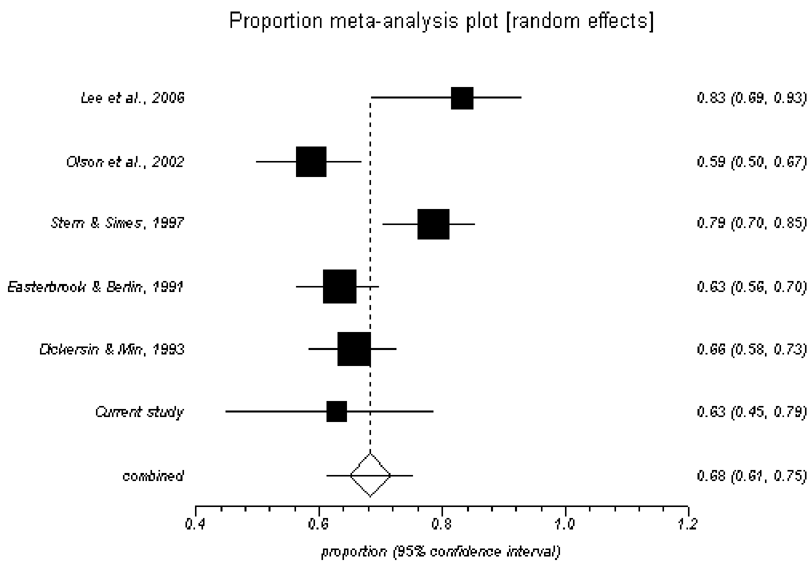

Figure 2: Forest plot of proportions of published, statistically positive articles in the medical research literature and the current study.

6 Discussion

Coming back to research questions 1 and 2, our study showed that 63% of the published computer science education articles had statistically positive results. Alone that proportion is hard to interpret; however, it takes on meaning when we compare it to similar studies from medical research. As Figure 2 showed, the proportion of published, statistically positive articles from the computer science education literature falls within the lower end of the expected range of proportions from medical research; in essence, there is no evidence that computer science education research articles and medical research articles differ in the proportion of published, statistically positive articles. The similarity of those proportions lead us to wonder about the degree to which the results of publication bias studies from other fields, especially medical research, can be generalized to the field of computer science education. If one were to generalize what is known about publication bias in the field of medicine to the field of computer science education, one could conclude also that in computer science education authorial publication bias is much more of a threat than editorial publication bias [Lee et al. 2006, Olson et al. 2002] and that

submitted manuscripts are more likely to be published if they have high methodological quality, a [randomized control trial] study design, descriptive or qualitative analytical methods and disclosure of any funding source, and if the corresponding author lives in the same country as that of the publishing journal. Larger sample size may also increase the chance of publication. [Lee et al. 2006, p. 621]

Coming back to research question 3, we found that the only statistically significant predictor of an article's having statistically positive results was the sex of the first author; the odds of an article's having statistically positive results was 12 times greater if the first author was a female. Neither whether an investigation was reported in a journal or conference proceeding nor whether the source of funding was disclosed were significant predictors of an article's having statistically positive results. Lee [Lee et al. 2006] found that gender of the first author was not a statistically significant predictor of editorial publication bias in medical articles. To our knowledge, this paper is the first piece of research that has investigated whether gender is a predictor of *authorial* publication bias.

Explanations for why articles with female first authors tended to have statistically positive results might have to do with the gender gap in computer science education. Female authors might have perceived that they needed to present only statistically positive results to establish what Irani [2004] calls "an identity of competence" in a male-dominated field. Similarly, our finding might be a manifestation of the phenomenon in computer science and engineering that "women tend to underestimate their abilities while men tend to overestimate theirs" [Irani 2004, p. 196]. Alternately, editors might have tended to accept males' statistically negative papers at a higher rate than they tended to accept females' negative papers. However, we do not believe this last explanation to be the case because the sex of the first author was not a predictor of editorial bias in the previous research and because reviews were reportedly done using double-blind reviewing. Nevertheless, given the fact that double-blind reviewing is not common in other areas of computer science (for example, IEEE Multimedia and Expo Conference submission requires full author names and affiliations), it would be interesting to evaluate the effects of gender of the first author in the areas of computer science where only single-blind reviewing methodology is used. Another explanation is simply that there was a sampling error or that there was a confound that we had not discovered. Replication is needed to determine which of these explanations, or combinations of explanations, is indeed the case.

Randolph et al. [2007] found that there were no statistically significant differences between computer science education journal articles and conference papers on five indicators of methodological quality–whether an experimental design was used, whether an explanatory descriptive (e.g., qualitative) design was used, whether attitudes were used as the sole dependent variable, whether the one-group posttest-only design was the only experimental design was used, and whether anecdotal evidence was the only type of evidence presented. Therefore, it is not surprising that there was not a statistically significant difference, either, in the percentages of statistically positive results between computer science education journal articles and conference papers. See Randolph [2007] for an in-depth discussion of the methodological quality of computer science education research articles.

We investigated whether having a source of funding disclosed might have been a predictor of having statistically positive results; we thought that there might be a tendency to please funders by making statistically positive results public and keeping negative results private. We take it as a good sign for the field that we found no evidence that such a tendency exists.

One unexpected finding was that we noticed that authors tended to emphasize statistically positive results by setting them off in tabular form or with mathematical symbols and to deemphasize statistically negative results by reporting them only in written text (without mathematical symbols). In fact, we estimate that only examining the inferential tests that are presented in tabular form or are set off with mathematical symbols would lead to a false positive in 10% of cases.

Another unexpected finding was that in all of the articles that we reviewed, none of the authors corrected for the number of pairwise inferential tests that had been conducted. For example, an author might conduct twenty t-tests and then report that a test was statistically significant because the p-value for that particular test was below .05. However, one would expect that one out of twenty t-tests would be significant given chance anyway, because a p-value of .05 indicates that a result that large or larger would only be found 5/100 times, or 1 in 20 times, given chance. [4] Had the authors of the articles we reviewed corrected for the number of inferential tests that had been conducted, the percentage of positive articles would have been much lower because the criterion for what is statistically significant result would have been much more conservative. Unfortunately, since this was an unplanned observation we do not have an estimate of how much lower the results would have been. We plan to examine this phenomenon in more detail in future research.

[4] The author could have corrected for the number of inferential tests conducted by applying the Bonferroni correction (see [Stevens 1999]). For fear of leading our readers astray, we have to report that there are, however, better ways of dealing with multiple comparisons than conducting all possible t-tests and applying a Bonferroni correction. One alternative is to use an omnibus ANOVA test followed by the Tukey procedure [Stevens 1999]. Another alternative is to use planned comparisons [Rosenthal et al. 2000].

7 Limitations and Summary

The primary limitation of our study was that editorial bias and authorial bias were confounded. Because we did not have access to submitted unpublished manuscripts, we could not determine the degree to which editorial bias uniquely influences the computer science education literature. Similarly, because we did not have access to every article that was written but not submitted, we could not determine the degree to which authorial bias uniquely influences the literature. This last limitation, however, is a limitation in all conventional publication bias studies. Nonetheless, we were able to estimate the proportions of statistically positive and statistically negative results within the computer science education research and compare those estimates to the estimates in the field of medical research, in which there is much information about publication bias. The fact that the proportions seemed to match across fields provides some preliminary evidence that publication bias findings from medical research might generalize to computer science education research, and vice versa. However, a body of evidence needs to be built before that generalization can be confidently made. Thus far, the evidence is purely circumstantial.

In conclusion, our results have shown that about 63% of published computer science education articles have statistically positive results, (b) that the percentage is very similar to the percentage in medical research, and (c) that articles with female first authors are much more likely to report statistically positive results than articles with male first authors. We also noted that (d) there is a tendency for authors to emphasize statistically positive findings by reporting them in tabular form or with mathematical symbols and to deemphasize statistically negative findings by presenting them in written text without mathematical symbols and noted that (e) authors do many pairwise comparisons but do not statistically correct for the amount of comparisons made.

It is our hope that this article will help shape and move forward the debate about publication bias in the computer science education literature. We also hope that the information presented here will be put to good use by the creators, consumers, funders, and gatekeepers of computer science education research. For example, we hope that this information will encourage authors to not withhold manuscripts that have negative results and to report negative and positive findings with equal care. We hope that editors and reviewers will give as much consideration to articles with positive results as to articles with negative or neutral results–both types of articles are equally necessary in building an accurate research record.

Acknowledgments

This research was supported in part by a Special Projects Grant from ACM SIGCSE. Justus J. Randolph is now at the Faculty of Education, University of Lapland, Finland.

References

[Agresti 1996] Agresti, A.: "An introduction to categorical data analysis"; New York: Wiley & Sons (1996).

[Brennan and Prediger 1981] Brennan, R. L., Prediger, D. J.: "Coefficient kappa: Some uses, misuses, and alternatives"; Educational and Psychological Measurement, 41, 687-699 (1981).

[Cochrane Bias Methods Group n.d.] Cochrane Bias Methods Group: "Types of bias"; Retrieved February 21, 2008 from `http://www.chalmersresearch.com/bmg/types_bias.html`.

[Dickersin and Min 1993] Dickersin, K., Min, Y-I.: "NIH clinical trials and publication bias"; The Online Journal of Current Clinical Trials.Document number: 50 (1993).

[Easterbrook and Berlin 1991] Easterbrook, P. J., Berlin, J. A.: "Publication bias in clinical research"; Lancet, 336(8746), 867-893 (1991).

[Fincher and Petre 2004] Fincher, S., Petre, M. (Eds.): "Computer science education research"; London: Taylor & Francis (2004).

[Grimm 2006] Grimm, D.: "A cure for the common trial"; Science, 312, 835-837 (2006, May).

[Higgins and Thompson 2002] Higgins, J. P. T., Thompson, S. G.: "Quantifying heterogeneity in a meta-analysis"; Statistics in Medicine, 21, 1539-1558 (2002).

[Irani 2004] Irani, L.: "Understanding gender and confidence in CS course culture"; In Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education SIGCSE '04, (pp. 195-199) (2004).

[Lee et al. 2006] Lee, K. P., Boyd, A. E., Holroyd, J. M., Bacchetti, P., Bero, L. A.: "Predictors of publication: Characteristics of submitted manuscripts associated with acceptance at major biomedical journals"; Medical Journal of Australia, 184(12), 621-626 (2006).

[Margolis and Fisher 2001] Margolis, J., Fisher, A.: "Unlocking the clubhouse: Women in computer science"; Cambridge: MIT Press (2001).

[National Research Council 1994] National Research Council: "Academic careers for computer scientists and engineers"; Washington, DC: National Academy Press (1994)

[Olson et al. 2002] Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., Zhu, Q., Reiling, J., Pace, B.: "Publication bias in editorial decision making"; Journal of the American Medical Association, 287(21), 2825-2828 (2002).

[Prechelt 1997] Prechelt, L.: "Why we need an explicit forum for negative results"; Journal of Universal Computer Science, 3(9), 1074-1083 (1997).

[Randolph 2007] Randolph, J. J.: "Computer science education research at the crossroads: A methodological review of computer science education research: 2000-2005"; Unpublished doctoral dissertation, Utah State University, Utah. Retrieved January 16, 2008, from `http://www.archive.org/details/randolph_dissertation` (2007).

[Randolph et al. 2007] Randolph, J.J., Julnes, G., Bednarik, R., Sutinen, E.: "A comparison of the methodological quality of articles in computer science education journals and conference proceedings"; Computer Science Education, 17(4), 263-274 (2007)

[Riehl 2006] Riehl, C.: "Feeling better: A comparison of medical research and education research"; Educational Researcher, 35, 24-29 (2006).

[Riniolo 1997] Riniolo, T.C.: "Publication bias: A computer-assisted demonstration of excluding nonsignificant results from research interpretation"; Teaching of Psychology, 24(4), 279-282 (1997).

[Rosenthal et al. 2000] Rosenthal, R., Rosnow, R. L., Rubin, D. B.: "Contrasts and effect sizes in behavioral research: A correlational approach"; UK: Cambridge UP (2000).

[Selby et al. 1998] Selby, L., Ryba, K., Young, A.: "Women in computing: What does the data show?"; ACM SIGCSE Bulletin, 30(4), 62-67 (1998).

[StatsDirect 2005] StatsDirect Ltd.: "StatsDirect statistical software [Computer software and manual]"; England: StatsDirect Ltd. (2005).

[Stern and Simes 1997] Stern, J. M., Simes, R. J.: "Publication bias: Evidence of delayed publication in a cohort study of clinical research projects"; British Medical Journal (315), 640-645 (1997).

[Stevens 1999] Stevens, J.: "Intermediate statistics: A modern approach (2nd ed.)"; Mahwah, NJ: Earlbaum (1999).