Publication Bias Currently Makes an Accurate Estimate of the Benefits of Enrichment Programs Difficult: A Postmortem of Two Meta-Analyses Using Statistical Power Analysis

Russell T. Warne

Utah Valley University

Completed August 22, 2016

Russell T. Warne, Department of Behavioral Science, Utah Valley University.

The author appreciates the feedback of Alyce Odasso on a draft of this manuscript.

Correspondence concerning this paper should be addressed to Russell T. Warne, Department of Behavioral Science, Utah Valley University, 800 W. University Parkway MC 115, Orem, UT 84058. rwarne@uvu.edu

Abstract

Recently Kim (2016) published a meta-analysis on the effects of enrichment programs for gifted students. She found that these programs produced substantial effects for academic achievement (g = .96) and socioemotional outcomes (g = .55). However, given current theory and empirical research these estimates of the benefits of enrichment programs are unrealistically high. In this manuscript I make the argument that Kim's results—and those from a previous meta-analysis on the same topic (Vaughn, Feldhusen, & Asher, 1991)—are likely distorted by publication bias. Using statistical power analysis, Schimmack's (2012) incredibility index, and an examination of the correlation between sample size and effect size, I present circumstantial evidence that publication bias is distorting the enrichment literature. As a result, gifted education scholars and practitioners do not really know the effectiveness of enrichment programs. I conclude the manuscript by discussing how to reduce publication bias in the gifted education literature.

Publication Bias Currently Makes an Accurate Estimate of the Benefits of Enrichment Programs Difficult: A Postmortem of Two Meta-Analyses Using Statistical Power Analysis

One of the definitions for the word "incredible" in *Oxford English Dictionary* is ". . . of a degree beyond what one would *a priori* have conceived as possible; inconceivable, exceedingly great." Gifted education scholars and practitioners often see incredible accomplishments among the people they study. Whether it is impressive performance on psychometric tests (e.g., Gross, 2004), the amazing talents of child prodigies (e.g., Ruthsatz, Ruthsatz, & Ruthsatz Stephens, 2014), or the stunning careers that some children with high abilities have in adulthood (e.g., Lubinski, Benbow, & Kell, 2014), the word "incredible" suits the accomplishments and abilities of many of the people that gifted education specialists work with. Observing these incredible individuals is one of the reasons gifted education is such an exciting field to work in.

But there is another definition of "incredible" that occasionally applies to work in gifted education: "Not credible: that cannot be believed; beyond belief." This definition of "incredible" came to my mind when I read a recent meta-analysis by Kim (2016) on the effectiveness of enrichment programs for gifted students. According to Kim, enrichment give gifted students ". . . richer and more varied content through modification and supplementation of content in addition to standard content in the regular classroom . . ." (2016, p. 103). Kim's results are incredible in this sense because she found that the average effect size (Hedges's *g*) for the academic outcomes of these add-on enrichment programs for gifted students was .96. For socioemotional outcomes the effect size was .55.

Why I Found Kim's Meta-Analysis Surprising

Past Empirical Data

Although these numbers are not so high as to be inherently absurd, these effect size estimates are much higher than what is normally found in gifted education research. A suitable comparison is Rogers's (2007) meta-analyses of various acceleration and ability grouping strategies. Of 23 grouping interventions in Rogers's report, only one—grade skipping—has a stronger effect size (Cohen's d = 1.00) for academic outcomes than what Kim (2016) found for enrichment programs. Of the 14 interventions that had enough research on social or self-esteem outcomes none was higher than d = .47 (Rogers, 2007, p. 388). Taken at face value this would indicate that enrichment provides more socioemotional and academic benefits for gifted children than *any* ability grouping or acceleration intervention, except grade skipping.

What surprised me about Kim's (2016) results was that they seem contradictory in light of much of the empirical research and theory on gifted education and human learning, even though her procedures were sound. For example, these results are outliers compared to previous meta-analyses, such Kulik and Kulik's (1992, pp. 75-76) finding (as part of a larger meta-analysis on grouping programs) that the effect size for academic outcomes of enrichment programs was .41. For socioemotional outcomes, the effect size in Kulik and Kulik's (1992) study was merely .10. Another early meta-analysis on enrichment programs (Vaughn, Feldhusen, & Asher, 1991) produced an estimate of d = .65 for the academic benefits of enrichment programs and d = .11 for self-concept.

Individual studies on enrichment programs also make Kim's (2016) meta-analysis appear supprising. For example, in one of the largest investigations of an enrichment program Callahan and her colleagues investigated the academic impact of a poetry enrichment curriculum and a research enrichment curriculum on three cohorts totaling 2,905 students. The median effect size (as quantified by Spybrook's δ, a standardized mean difference effect size analogous to Cohen's

d that is appropriate for hierarchical linear models) was .67 (Callahan, Moon, Oh, Azano, & Hailey, 2015). This effect size is noticeably smaller than the .96 effect size in Kim's (2016) meta-analysis, even though Callahan et al.'s (2015) enrichment programs had a very high quality curriculum, high treatment fidelity in implementation, and outcomes measured with custom-designed instruments closely aligned with the curriculum (Cheung & Slavin, 2016). These characteristics likely make the Callahan et al. (2015) enrichment program more effective than the typical school district enrichment program. And yet, this study did not have an effect size as strong as the mean effect size in Kim's (2016) meta-analysis.

Theoretical Concerns

From a theoretical standpoint, leading scholars of gifted education have been critical of enrichment. Some experts have discussed the concept of "educational dose," which refers to the strength of an educational intervention for gifted students (Gallagher, 2000; Wai, 2015; Wai, Lubinski, Benbow, & Steiger, 2010; Warne, 2016). These authors argue that for an intervention to be effective for gifted students it must be robustly different from the regular educational curriculum and administered for a long enough period of time to have an impact. Indeed, in some respects, Kim's (2016) results are consistent with the concept of the methodological dose, such as when she found that, "Studies with extremely high effect sizes were more intensive programs than other studies . . . more intensive enrichment programs influence more academic achievement of gifted students" (p. 108). However, the typical enrichment program is a pullout program that a gifted child participates in for no more than four hours per week (Callahan, Moon, & Oh, 2014, p. 7). Given the concept of an education dose, it is hard to imagine how the average enrichment program could have nearly the same educational impact as grade skipping, which is an intense intervention that operates every minute a child is in school.

Other experts see the needs of gifted children in a similar context and justify the existence of gifted programs or interventions on the basis of an educational need for a child (e.g., Ruf, 2005; Stanley, 1977). In one particularly scathing criticism of pull-out enrichment programs, VanTassel-Baska (1987) explained how curriculum differentiation and adjustment to educational programs is a full-time need for gifted children—which by definition cannot be met in a program in which a child spends the minority of their school time. Again, given the mismatch between full-time need and part-time intervention, the .96 effect size that Kim (2016) found seems unrealistically high when compared to treatments that occupy a larger proportion of a gifted child's school day, such as full-time ability grouping (d = .49), curriculum compacting (d = .83), or grade telescoping (d = .45; all effect sizes are for academic outcomes in Rogers, 2007, p. 388).

Although the theorizing is not as strong for the socioemotional impact of enrichment programs, it likewise seems difficult to explain why a less intense enrichment program would have a stronger impact than immersive programs, like Talent Searches (see Matthews, 2008, for a description of the social and emotional benefits of Talent Search programs). The strong effect size for socioemotional outcomes in Kim's (2016) meta-analysis is also difficult to explain when one considers that enrichment programs are not designed primarily to improve socioemotional outcomes.

How Did Kim Obtain Such Strong Effect Sizes?

Possibility 1: Errors in the Meta-Analysis

Some readers may ponder the strong effect sizes in Kim's (2016) meta-analysis and assume that her results are a consequence of problems with her work. I do not agree with this interpretation. Although Kim's (2016) results do not seem realistic in light of prior empirical

research or theory, Kim did not commit any methodological errors in her meta-analysis. Indeed, she followed modern standards for meta-analysis, such as the Meta-Analysis Reporting Standards (American Psychological Association, 2010), and her methodology conforms to guidelines from experts on how to conduct a meta-analysis (e.g., Cooper, 2010; Lipsey & Wilson, 2001; Steenbergen-Hu & Olszewski-Kubilius, 2016). For example, her search procedures were thorough, and she made an honest attempt to find all the grey literature (i.e., studies that were not published in peer reviewed journals) that she could. The transparency of Kim's work and her careful efforts to follow best practice make the possibility of an error on her part extremely unlikely.[1]

Possibility 2: Problems with the Data

If Kim performed her meta-analysis correctly, then her unrealistic results did not arise from her procedures. Instead, the results must originate with the studies in her meta-analysis. Although it is not completely clear what is wrong with the studies that Kim meta-analyzed, I argue in this manuscript that the problem may be publication bias. Publication bias occurs when studies that retain the null hypothesis are less likely to be submitted or accepted for publication (Greenwald, 1975). Publication bias can also manifest itself as a bias against publishing replications because gatekeepers (i.e., journal editors and reviewers) reject a study for reasons unrelated to its quality, such rejecting studies because they do not present new knowledge (Makel & Plucker, 2014, 2015). This bias against null findings and replications has been observed in many fields, including medicine (Coronado-Montoya et al., 2016), sociology (Gerber & Malhotra, 2008), psychology (Laws, 2013), and education (Makel & Plucker, 2014). These unpublished studies tend to have effect sizes closer to zero than published studies. As a result, the published studies on a topic generally present stronger effect sizes, and meta-analyses

on these studies make phenomena appear more robust and interventions more effective than they really are (Cheung & Slavin, 2016). Thus, publication bias is a plausible explanation for Kim's (2016) optimistic results of the impacts of enrichment programs for gifted students.

Purpose of This Manuscript

Given the mismatch between Kim's (2016) results and those of other meta-analyses and of gifted education theory, I chose to investigate the possibility of publication bias in Kim's (2016) meta-analysis and a similar, earlier analysis by Vaughn et al. (1991). Although there is no "smoking gun" evidence of severe publication bias in the enrichment literature, I found four pieces of circumstantial evidence that suggest the presence of some degree of publication bias. Three of these pieces of evidence are based on statistical power analysis—a topic that will be explained in the next section. In this manuscript, I will also explain how to evaluate publication bias, demonstrate four procedures that can be used with smaller bodies of research, and discuss the likelihood of publication bias in the enrichment literature. I will also close the manuscript by making suggestions for changes in the culture of gifted education scholarship that would reduce the prevalence of publication bias.

There are three benefits to performing statistical power analysis on these meta-analyses. First, statistical power analysis simply and intuitively demonstrates the possible presence of publication bias. Second, this manuscript will show that the possible presence of publication bias should make educators and scholars cautious about interpreting the actual impact of these programs. Finally, statistical power is an important concept in its own right that has been neglected in psychological and educational research for too long. Through this study I hope that gifted education scholars will be more cognizant of statistical power as they plan their studies and evaluate the work of others. I believe that this article will move readers beyond the normal

preaching about the importance of statistical power (e.g., Cohen, 1992) or the dangers or publication bias (e.g., Greenwald, 1975) by showing a real-world example where publication bias may endanger an accurate understanding of the impact of an intervention.

Evaluating Publication Bias

There are several ways to evaluate publication bias. One recent development is the examination of distributions of *p*-values to search for a surplus of *p*-values below .05 (Ginsel, Aggarwal, Xuan, & Harris, 2015; Masicampo & Lalande, 2012). Other researchers (e.g., Flore & Wicherts, 2015), examine the correlation between sample size and effect size, which should be zero without publication bias and negative if publication bias is present. Another popular method is to use a funnel plot in a meta-analysis to search for asymmetry in the distribution of effect sizes (e.g., Roth et al., 2015). Other methods of investigating publication bias are available (see Nuijten, van Assen, Veldkamp, & Wicherts, 2015; Steenbergen-Hu & Olzewski-Kubilius, 2016).

Much to her credit, Kim (2016) investigated the possibility of publication bias in the enrichment literature using the trim-and-fill method. This standard procedure removes extreme, small studies from a meta-analysis (this is the "trim") and imputes missing—usually smaller—effect sizes (the "fill"), and recalculates the new effect size. If publication bias is not a major problem, then the new effect size after the trim-and-fill method should be similar to the meta-analysis's effect size (see te Nijenhuis, Willigers, Dragt, & van der Flier, 2016, p. 124, for a brief, accessible explanation of this method). When Kim (2016) performed the trim-and-fill procedure, she found that the imputed effect size actually increased slightly. This would generally indicate that publication bias was not a problem in a meta-analysis, and Kim interpreted it as such (see p. 108).

Kim's (2016) test for publication bias undermines my claim that publication bias distorted her results. However, the trim-and-fill method requires a large number of effect sizes to effectively detect publication bias (Nuijten et al., 2015). With only 13 effect sizes in her meta-analysis of academic outcomes and 18 effect sizes in her meta-analysis of socioemotional outcomes, Kim (2016) did not have the statistical power to detect publication bias in the enrichment literature. Again, this is not her fault. It is a deficiency in the research base that there are so few studies on the consequences of enrichment programs.

Statistical Power as a Tool for Investigating Publication Bias

Because the trim-and-fill method was inadequate to detect publication bias in the enrichment literature, other methods should be used that can detect publication bias in a relatively small number of studies. Schimmack (2012) proposed such a method based on statistical power.

What is statistical power? Statistical power is the likelihood of a study to reject a false null hypothesis. Studies with high statistical power have a high probability of rejecting a false null hypothesis, while rejecting the null hypothesis is unlikely when a study has low statistical power—even if the actual effect is strong. Statistical power is the product of four aspects of a study: (a) the alpha value of the null hypothesis test, (b) the magnitude of the effect size, (c) the sample size, and (d) the study design. The most flexible of these four study characteristics is the alpha value of the null hypothesis because alpha is completely arbitrary. However, given the ubiquity of the default .05 alpha value, raising alpha is problematic and may in some instances appear to be an effort to manipulate the study's results to produce a desired outcome. Aspect (b), the magnitude of the effect size, is largely out of the control of researchers.

This leaves adjusting sample size and study design as the typical methods of increasing statistical power. If all other aspects are equal, then studies with larger sample sizes have larger power because larger sample sizes make it easier for all statistical tests to reject the null hypothesis (Cohen, 1994; Rodgers, 2010; Sedlmeier & Gigerenzer, 1989; Thompson, 1992). In regards to study design, usually within-subjects designs have higher statistical power than between-subjects designs (Zimmerman, 1997). If a researcher does choose a between-subjects design—which is common in gifted education research—then having groups that are balanced in size also increases power (Kline, 2013).[2] Generally, researchers recommend that statistical power for a study be at least .80 (Cohen, 1962), though higher power is always desirable.

Why statistical power matters. Statistical power is important because publication bias means that studies with low power are less likely to reject the null hypothesis and therefore less likely to be published. Therefore, low statistical power contributes to a distorted scientific literature in many fields. Paradoxically, studies with low statistical power are not inevitably hidden away. Because of sampling error and Type I error, some studies with low power do indeed reject the null hypothesis. These studies capitalize on chance and sampling error—often unbeknownst to the author—and are less likely to replicate in the future. Yet, they are more likely to be published anyway because the null hypothesis was rejected. As a result, the studies with low power that are published often seem inconsistent and contradictory (Maxwell, 2004; Schimmack, 2012).

Ironically, this means that low statistical power—when combined with publication bias—results in "the worst of both worlds." Studies that fail to reject the null hypothesis languish unpublished—often because of low power—and the studies that are published often are not replicable. Therefore, the research literature becomes full of unstable findings that are

nongeneralizable, unreplicable, and difficult to build a theory around. The distorting effects of publication bias and low statistical power are not just some theoretical quirk of methodology. There is strong evidence to indicate that the publication bias on psychological topics like stereotype threat (Ganley et al., 2013), mindfulness treatments (Coronado-Montoya et al., 2016), and social priming (Vadillo, Hardwicke, & Shanks, 2016) is so strong that the true effect sizes for these phenomena may be zero. This could be why the research on these topics is so confusing and contradictory. The inflated effect sizes in the research on these topics makes planning future studies difficult because using inflated effect sizes to estimate statistical power for a study will result in inflated statistical power—which will not accurately reflect the true probability of rejecting a null hypothesis.

Using Statistical Power to Investigate Publication Bias

Schimmack (2012) proposed a method of investigating publication bias via statistical power. His method is based on the fact that if no publication bias is present, then the average statistical power of a set of null hypothesis tests (expressed in statistical notation as $1 - \beta$) will be equal to the proportion of tests that reject the null hypothesis (which I abbreviate as $prop_{reject}$). Schimmack (2012) created the Incredibility Index (IC) to quantify the degree of discrepancy between these two values:

$$IC = prop_{reject} - (1 - \beta)$$

If publication bias favors studies that reject the null hypothesis, then IC is positive. If IC is negative, then studies that retain the null hypothesis are more likely to appear in the literature. Values close to zero indicate that there is little, if any, publication bias.

Although the logic of Schimmack's (2012) Incredibility Index is sound, the mean statistical power of a set of studies ($1 - \beta$) is a positively biased estimate of the true statistical

power of the studies (Yuan & Maxwell, 2005). Therefore, in this manuscript I will calculate IC values on the basis of the median power of a set of studies, rather than the mean power.

Another method in which $prop_{reject}$ and mean statistical power can be used to investigate population bias is to use the overall (i.e., median) statistical power to create a binominal distribution of expected numbers of rejected and retained null hypotheses under the assumption that there is no publication bias. This value can then be compared to $prop_{reject}$ in either Fisher's exact test or in a chi-square goodness-of-fit test with one degree of freedom (Ioannidis & Trikalinos, 2007). If the null hypothesis is retained, then there is no evidence of publication bias. A rejected null hypothesis, however, would suggest the presence of publication bias.

Methods

The first step in using statistical power to estimate publication bias is to calculate the power for every effect size in each meta-analysis using the procedures explained in Cohen (1988). This required converting all of the Hedges's *g* effect sizes in Kim's (2016) meta-analysis into Cohen's *d* values. A formula provided by Durlak (2009, p. 928) was used for this purpose. Because Hedges's *g* and Cohen's *d* are both standardized mean differences, these values are very similar, though there are slight discrepancies. Additionally, all of the effect sizes in the Vaughn et al. (1991) meta-analysis were recalculated after consulting the original studies because of discrepancies arising from contradictory or confusing information in Vaughn et al.'s (1991) report. Thus, the effect size values reported in this manuscript differ from those in the previous meta-analyses.

After all effect sizes were converted to Cohen's *d* values, a statistical power value was calculated for each of four different effect sizes: (a) the study's observed effect size, (b) the mean effect size reported by the meta-analyst for the group, (c) an effect size of $d = .50$, and (d) the

median effect size reported by Rogers (2007) for ability grouping and acceleration interventions for either academic outcomes (d = .45) or socioemotional outcomes (d = .15). The only difference among these power values was in the effect size used to calculate them.

Using multiple effect sizes is best practice for examining statistical power in a meta-analysis (Ioannidis & Trikalinos, 2007), and these effect sizes will produce differing estimates of statistical power. Using the study's observed effect size (i.e., option a) in statistical power calculations is called *post hoc* power. *Post hoc* power is simply a function of a study's *p*-value (Hoenig & Heisey, 2001) and is inflated when publication bias is present, especially when sample sizes are small (Schimmack, 2012). Statistical power calculations based on option (b) also is inflated because it is a function of *post hoc* power from the meta-analysis's constituent studies. Therefore, I presented these statistical power calculations here for comparison purposes only.

Statistical power calculations based on option (c) have a history that dates to Cohen's (1962) study of statistical power in leading psychology journals. In that article he stated that an effect size of d = .50 had a "medium" strength (a benchmark he carried into his landmark 1988 book) and used that threshold to find that the mean statistical power in psychology was .48. Others (e.g., Osborne, 2008) have used the same standard to estimate power in the social sciences, including in educational research. I used this value to ensure that my statistical power calculations were comparable to similar calculations of power. Finally, statistical power calculations based on Rogers's (2007) median effect sizes for acceleration and ability grouping interventions were used because these were empirically derived from gifted education research. Because the effect sizes in options (c) and (d) were generally smaller than (a) and (b) in the enrichment literature, the statistical power estimates corresponding to effect sizes (c) and (d)

were usually lower. Having a range of power estimates is helpful; any one particular value may be inaccurate because it is impossible to know beforehand truly whether publication bias is present in a meta-analysis.

After the four estimates of statistical power have been calculated, an IC value was calculated for each article, along with an average IC value across the entire collection of studies. I then conducted a chi-square goodness-of-fit test in order to determine whether there was a statistically significant difference between the number of effect sizes that rejected the null hypothesis and the expected number of rejected null hypotheses if there were no publication bias. To provide another source of evidence about publication bias I also calculated the correlation between sample size and effect size within each group of studies, which should be zero if no publication bias is present. This latter procedure is the only examination of publication bias in the article that is not based in some way on statistical power.

Results

Description of Effect Sizes

All effect sizes from both meta-analyses were compared in two groups: effect sizes measuring academic achievement outcomes, and effect sizes measuring all other outcomes. The former group consisted of 15 effect sizes. The latter group consisted of 22 effect sizes and included socioemotional and critical thinking outcomes, plus one creativity outcome (from Kolloff, 1983) reported in the Vaugh et al. (1991) meta-analysis. All effect sizes are listed in Kim (2016) and Vaughn et al. (1991). Some of the effect sizes from the latter study were eliminated before analysis because statistical power could not be calculated, either because the original study was unpublished and now unavailable or because there was not enough information to estimate statistical power. Two studies were in both meta-analyses (Aldrich &

Mills, 1989; Feldhusen, Sayler, Nielsen, & Kolloff, 1990). For both of these studies I converted Kim's (2016) effect sizes to Cohen's *d* for purposes of my analyses.

Statistical Power and IC Estimates

The effect sizes for academic achievement outcomes are reported in Table 1. Across the two meta-analyses the effect sizes ranged from d = -0.10 to 2.99, with a median of 0.78. The four power estimates for each study are shown in the table. The median value for the four statistical power estimates are (a) 0.9424 for *post hoc* statistical power, (b) 0.9701 for the mean effect size from the meta-analyses, (c) 0.5171 for a "medium" effect size, and (d) 0.4376 for the median effect size from Rogers's (2007) report. The corresponding IC values are (a) 0.0464, (b) -0.1110, (c) 0.3659, and (d) 0.4374.

A powerful influence on statistical power is the choice of a between-subjects or within-subjects designs, with the latter almost always having higher statistical power (Zimmerman, 1997). However, within-subjects designs are slightly more vulnerable to several threats to internal validity, such as history events and maturation (see Winch & Campbell, 1969). As a result, it is likely possible that these effect sizes (and therefore statistical power estimates) are inflated. Therefore, the table also displays the average statistical power and IC values for only the between-subjects designs. Comparing the statistical power and IC values for the subset of between-subjects designs with the entire set of effect sizes from Table 1 shows that the subset has marginally lower power and corresponding higher IC values.

INSERT TABLES 1 & 2 ABOUT HERE.

Table 2 shows the 22 effect sizes for other outcomes, which ranged from d = -0.36 to 13.42, with a median of 0.46. The four power estimates for each study are shown in the table. The average statistical value for the four statistical power estimates are (a) 0.6541 for *post hoc*

statistical power, (b) 0.9945 for the mean effect size from the meta-analyses, (c) 0.9870 for a "medium" effect size, and (d) 0.2849 for the median effect size from Rogers's (2007) report. The corresponding IC values are (a) -0.1541, (b) -0.4945, (c) -0.4870, and (d) 0.2151. Again, for the non-academic outcomes the studies with a within-subjects design were removed and the power estimates were recalculated. Just as in Table 1, the removal of the within-subjects designs lowered all of the statistical power estimates, though the IC values did not change in a consistent fashion.

Chi-Square Tests

For the academic achievement effect sizes the median statistical power for a "medium" effect size of $d = .50$ was .5171. Therefore, it would be expected that 51.71% of effect sizes would be statistically significant. Eleven of these 15 effect sizes (73.33%) did reject the null hypothesis. A chi-squared goodness of fit test was conducted to test whether the observed number of rejected null hypotheses differed from the expected value, given the statistical power. This discrepancy was not statistically significant ($\chi^2 = 2.809$, df = 1, $p = .094$). Using the statistical power estimate calculated on the basis of Rogers's (2007) report, the discrepancy between the expected number of rejected null hypotheses (43.76%) and the actual number of rejected null hypotheses (73.33%) was statistically significant ($\chi^2 = 5.331$, df = 1, $p = .021$). When the within-subjects effect sizes were removed, there were more rejected null hypotheses than expected for the academic achievement effect sizes, regardless of whether statistical power was calculated on the basis of a "medium" effect size of $d = .50$ ($\chi^2 = 5.944$, df = 1, $p = .015$) or the median effect size of $d = .45$ from Rogers's (2007) article ($\chi^2 = 8.925$, df = 1, $p = .003$).

For the non-academic outcomes there were 22 effect sizes, 11 (50%) of which rejected the null hypothesis. The median statistical power for an effect size of $d = .50$ was .9870; for

Rogers's (2007) median effect size of .15, the mean statistical power was .2849. For the first power estimate the differences between the number of actually rejected null hypotheses and the expected number was statistically significant (χ^2 = 1,653.88, df = 1, p < .001). These statistically significant results occurred because the number of rejected null hypotheses was *less* than expected given the statistical power. (Hence, the strong negative IC value of -0.4870.) When the basis of the expected number of rejected null hypotheses was calculated via the smaller statistical power value of .2849, then the results were also statistically significant (χ^2 = 6.070, df = 1, p = .014). However, this was due to a larger number of rejected null hypotheses than would be expected. Once the within-subjects studies were removed from analysis the discrepancy in the number of rejected null hypotheses was barely statistically significant for the statistical power calculated on the basis of an effect size of d = .50 (χ^2 = 3.967, df = 1, p = .046), though not for the median effect size of d = .15 from Rogers's (2007) article (χ^2 = 1.995, df = 1, p = .158).

Supplemental Test for Publication Bias

The final test for publication bias in this manuscript is to determine the correlation between the sample size and effect size within each table. For the academic achievement effect sizes this correlation was r = -.162 (p = .252) or ρ = -.131 (p = .664). These results did not substantially change when within-subjects designs were removed (r = -.171, p = .612; ρ = -.009, p = .979). For the other effect sizes, this correlation was r = -.151 (p = .224) or ρ = -.335 (p = .127). When the within-subjects designs were removed, the results changed inconsistently: r = -.177 (p = .201) or ρ = -.516 (p = .104).

Discussion

There is no perfect test of publication bias when examining a body of literature, and best practice is to use multiple tests in combination with one another (Vadillo et al., 2016).

Unfortunately, many of these tests are insensitive to publication bias when there are few studies in the meta-analyses. The trim-and-fill procedure that Kim (2016) performed is one such test that often cannot detect publication bias when the number of effect sizes is small. However, other procedures can be effective when there are a small number of effect sizes (e.g., Schimmack, 2012). When using two additional procedures based on statistical power, there were some conflicting results, but some evidence of publication bias emerged. This could explain the surprisingly large effect sizes in the enrichment literature.

Testing for Publication Bias With Statistical Power

In this manuscript I made four calculations of statistical power based on varying plausible effect sizes for the magnitude of the impact of enrichment programs on gifted students. These power estimates and $prop_{reject}$ were used to calculate IC for each statistical power scenario. Statistical power was highest when based on a study's observed effect size (i.e., *post hoc* power) and for the effect size averages in the Kim (2016) and Vaughn et al. (1991) meta-analyses. On the surface, this may seem like an indication that publication bias is completely absent from the enrichment literature. However, this is an example of begging the question. Both of these statistical power estimates—and therefore the IC values based on them—are based on the assumption that there is no publication bias. Therefore, these values cannot be used to test for publication bias because the absence of publication bias is already taken for granted.

Because the traditional "medium" effect size in power calculation (i.e., $d = .50$) and the median effect size in Rogers's (2007) report of the effectiveness of acceleration and ability grouping interventions for gifted children (i.e., $d = .45$) were nearly equal, statistical power estimates and IC values were similar. The statistical power for academic achievement effect sizes were .5171 and .4376, and IC values were positive (.2162 and .2957). These statistical power

estimates indicate that approximately half of studies in the enrichment literature were not capable of detecting a "medium" sized effect. It is noteworthy that these estimates are similar to what Cohen (1962) found over half a century ago when he calculated the mean statistical power of the abnormal and social psychology literature as .48. The corresponding IC values indicate that a larger percentage of studies rejected the null hypothesis of $d = 0$ than would be expected, providing modest evidence of publication bias. The chi-square test supports this modest evidence of publication bias, with the test based on the higher effect size (i.e., $d = .50$) showing no publication bias, but the test based on the slightly smaller effect size of $d = .45$ indicating publication bias. The ability of this chi-squared test to show publication bias is inhibited by the small number of effect sizes (15).

It is important to note, though, that some of these results are driven by the inclusion of within-subjects designs in the analyses. When only between-subjects designs were investigated, the evidence of publication bias strengthened. Statistical power decreased (to .4523 and .3808) and IC values approximately rose (to .3659 and .4374). Additionally, even though there were now fewer effect sizes, the chi-squared tests *were* statistically significant, indicating that more studies were rejecting the null hypothesis than would be expected with a population effect size of $d = .50$ or $d = .45$ and the total absence of publication bias. In other words, the large proportion of studies that show benefits for enrichment programs are literally "too good to be true" and would be unlikely if there were no publication bias. Supporting the claim of publication bias in the enrichment literature is the negative correlation between sample size and effect size, though these results were not statistically significant, perhaps because of the small number of effect sizes.

When the non-academic achievement effect sizes were tested for publication bias, the results were more inconsistent, likely because the two relevant effect sizes used to calculate statistical power varied more (i.e., d = .50 and d = .15). The IC values for these effect size estimates provided inconsistent results. For a population effect size of d = .50, the IC was -.4870—indicating that (if anything) there was a publication bias against statistically significant results. Although this is not unheard of (e.g., Roth et al., 2015), this is surprising. On the other hand, for a population effect size of d = .15 (based on Rogers's, 2007, report), the IC was .2151, indicating weak evidence of modest publication bias. The removal of within-subjects designs did not substantially change these results, though statistical power did decrease greatly.

Like in the academic effect sizes, when comparing the number of actual and expected rejected null hypotheses, given the statistical power estimates, the results for non-academic achievement effect sizes were contradictory. Using a "medium" population effect size estimate of d = .50, the results—again—indicated a publication bias in favor of the null hypothesis. However, for an effect size of d = .15, the chi-square test indicated a publication bias against the null hypothesis. When the within-subjects designs were removed from analysis, these results change substantially: the statistical test based on a population effect size of d = .50 indicated publication bias against the null hypothesis, while the chi-squared test based on a population effect size of d = .15 did not indicate any publication bias.

Supplemental Test of Statistical Power

The other test for publication bias that I performed was a calculating a correlation between the effect size and sample size. Although none of the correlations were statistically significant (probably as a result of the small number of effect sizes), it is noteworthy that all were negative, ranging from -.009 to -.516. This is circumstantial evidence for publication bias against

the null hypothesis; if there were no publication bias, an approximately equal number of correlations would be positive and negative.

Is there Publication Bias in the Enrichment Literature?

The difficulty of testing for publication bias is that the researcher uses published studies to estimate the presence and size of the unknown, unobservable body of research. Indeed, it is difficult to state confidently (a) whether publication bias exists in a field and (b) the size of the unpublished corpus of studies unless publication bias is extremely strong (e.g., Vadillo et al., 2016) or the body of literature is large and tailored around a very specific research question (e.g., Flore & Wicherts, 2015; Ganley et al., 2013). Neither of these situations applies to the literature on the effectiveness of enrichment programs for gifted children.

Nevertheless, there are tantalizing clues that indicate that publication bias is a problem in the enrichment literature. First, statistical power is low—especially for the more internally valid between-groups design—indicating that some studies that are published may be flukes that perhaps capitalize on Type I error to obtain statistical significance, and therefore publication. Second, IC values tend to be positive for the enrichment literature, indicating that more studies are published that reject the null hypothesis than one would expect without publication bias. This is especially true for studies where the dependent variable was academic achievement; among these studies the IC value reached a value as high as .4374—indicating that the percentage of studies that reject the null hypothesis may be up to 43.74 percentage points too high. These results generally are supported by the chi-square tests, though this support is often ambiguous because of—ironically—low statistical power arising from the small number of effect sizes. Finally, there are the negative correlations between sample size and effect size—an indication that smaller studies with less statistical precision dominate the enrichment literature. This seems

especially true when considering that the mean *n* was 71.3 (median *n* = 54), and only a single study (McCoach, Gubbins, Foreman, Rubenstein, & Rambo-Hernandez, 2014) had an *n* larger than 100.

Do Meta-Analyses Accurately Estimate the Benefits of Enrichment Programs?

All this talk of publication bias is irrelevant if a meta-analysis still provides accurate estimates of population effect sizes. Therefore, a skeptic may state that with a large effect size of *g* = .96 (in Kim, 2016, meta-analysis) or *d* = .65 (in Vaughn et al., 1991) that surely enrichment programs must be beneficial. After all, pooling together these studies produces a total sample size of 1,069 for the academic effect sizes and 4,767 for the non-academic effect sizes. I imagine that some readers would think that these *n*'s are large enough to produce accurate estimates of population effect sizes.

However, this line of thinking is fallacious. When a publication bias exists against retaining the null hypothesis—even in a small amount—the result is inflated effect sizes. Nuijten et al. (2015) called this the "replication paradox" (p. 172), wherein adding additional studies to the research literature actually makes population effect size estimates *less* accurate when there is a publication bias against retaining the null hypothesis. This effect is actually strongest when the population effect size and the average study size are both small. Both of these conditions are plausible characteristics of the enrichment literature—especially for academic achievement effect sizes. Therefore, the strong effect sizes in Kim (2016) and Vaughn et al.'s (1991) meta-analyses are no indication that enrichment programs are effective interventions.

Another possible argument against the presence of publication bias would be that the large number of studies showing the benefits of enrichment programs for gifted children cannot all be wrong. After all, the tables in this manuscript show that a majority (22 of 37) effect sizes

indicate a benefit for gifted children participating in enrichment programs. And most of the remaining effect sizes are statistically equal to zero. In fact, only one (from Coleman & Fults, 1982) is negative and statistically significant. Therefore, a tally of the effect sizes seems to indicate a clear "victory" for positive studies (22), compared to neutral (14) or negative (1) studies.

This methodology is called "vote counting," and meta-analysis experts have long recognized that it is an inherently flawed methodology that produces inaccurate estimations of the impact of treatments (Glass, 1976; Meehl, 1990), especially when the sample sizes in the body of literature are small (Hedges & Olkin, 1980; Schimmack, 2012; Vadillo et al., 2016). In fact, one of the reasons that social scientists accepted meta-analyses so quickly was because a few prominent early meta-analyses showed that vote counting was innately unsound (e.g., Schmidt & Hunter, 1977; Smith & Glass, 1977). Given the small sample sizes of studies of enrichment programs (especially for academic achievement outcomes), vote counting cannot provide accurate view of the literature.

Conversely, a skeptic could point to the 14 effect sizes that were statistically equal to zero as evidence that there is no publication bias against the null hypothesis in the enrichment literature. It seems surprising that over one-third of effect sizes would be statistically equal to zero if there were really bias against the null hypothesis. However, this ignores the fact that meta-analyses simplify the original articles they are based on. For example, in Carter's (1986) study, Vaughn et al. (1991) found that the effect size was a statistically insignificant $d = 0.24$, which I have verified independently. But the original study had three outcome measures. Carter used the highly statistically significant difference in one outcome—higher level thinking scores—to argue that the enrichment program was effective. Thus, even though the effect size of

interest to Vaughn et al. (1991) was not statistically significant, publication bias in favor of rejecting the null hypothesis could still exist. The fact that Carter did not reject the particular null hypothesis that meta-analysts were interested in does not indicate an absence of publication bias.

Another argument that a skeptic could make against my analysis is that power analysis requires setting an effect size *a priori*. Anyone could make a claim about publication bias via power analysis if they just set the effect size in the statistical power calculation to be small enough.[3] This would, in turn, increase the IC and produce an apparent surplus of rejected null hypotheses in the chi-square analyses. Thus, to the skeptic statistical power appears to be tautological. However, this argument ignores the fact that there are good reasons to test statistical power at the effect sizes in Tables 1 and 2. The "medium" effect size of $d = .50$, for example, ". . . would be a fairly noticeable phenomenon . . ." (Cohen, 1962, p. 147) in everyday life. It is reasonable to expect a program to be defined as "providing benefits" if its results are "fairly noticeable," and it is not unrealistic to ask that studies be designed to detect such an effect. Additionally, the median effect sizes from Rogers's (2007) collection of effect sizes are not plucked randomly from the ether. Rather, these are effect sizes that—based on prior data from other gifted education interventions—are reasonable standards by which to judge the enrichment literature.

Finally, those who still take Kim's (2016) findings at face value still must explain how enrichment programs can provide such strong benefits for gifted children when other, apparently more intensive interventions (e.g., subject acceleration) are apparently less effective. The weaker educational dose of enrichment programs should correspond to weaker effect sizes—but this is not true when comparing Kim's results with other meta-analyses (e.g., Rogers, 2007; Kulik & Kulik, 1992). Even if the theory of educational dose is flawed, it is still up to those who view

enrichment as being highly effective to provide a coherent theoretical explanation of *why* enrichment is more effective than other interventions for gifted children.

This section is a longwinded answer to the question in the heading: do meta-analyses accurately estimate the benefits of enrichment programs? I believe the short answer is no. There are just too many clues that the enrichment literature is the product of publication bias. When publication bias is present, meta-analyses cannot provide an accurate representation of the impact of a treatment. Indeed, publication bias may grossly inflate the apparent effectiveness of that treatment. It also inflates the *a priori* statistical power estimates that future researchers may calculate if they use Kim's (2016) mean effect sizes as their estimates of population effect sizes, thus perpetuating the problem of underpowered studies in the research literature.

One astute reviewer of this manuscript argued that Kim's (2016) meta-analysis results are empirical data and that a proper course of action would be to change my beliefs to match the data. I admit that if Kim's results had not been so different from my preconceived notions of the effectiveness of enrichment programs, I would not have examined her data more closely.[4] However, my belief in the distorting effect of publication bias on the enrichment literature is not a blind faith. Rather, I also have empirical data—in the form of four "clues" described above—that supports my beliefs. If my investigations of publication bias had been fruitless or if only one of these tests indicated a distorted literature base, then I would change my beliefs on the effectiveness of enrichment programs. While it is clear that none of these data unambiguously indicate publication bias, they do provide circumstantial evidence of that publication bias could be distorting the enrichment literature. As a result, I believe that a healthy skepticism towards massive effect sizes showing the effectiveness of enrichment interventions is in order. Kim's (2016) results of $g = .96$ seem too high, but I think that it would be unjustified to say that the

effect size should be zero. As to the true value of the population effect sizes, I am agnostic, and I believe that the publication bias is probably too severe at this time to provide an accurate estimate.

Dealing With the Aftermath

If I am correct and meta-analyses cannot tell gifted education scholars and practitioners the magnitude of the benefits of enrichment programs, then the next step is to decide how to deal with gifted education research and practice in light of the existence of a distorted literature summarized through distorted meta-analyses and literature reviews. Previously, the course of action for handling publication bias was for the meta-analyst (and their readers) to blame the research literature and wash their hands of the problem by stating that the estimated population effect size was the best estimate available. I have—even recently—been guilty of this practice (see Slade & Warne, 2016; Warne, 2011). Within the past few years assessing publication bias has become an essential step in conducting a meta-analysis, as demonstrated by Kim's (2016) work. However, the mere assessment of publication bias is no longer a sufficient course of action when confronting publication bias. Researchers must take more proactive action on this issue, and I have a few recommendations for improvement.

First, gifted education scholars and practitioners should admit their ignorance about the impact of many of their educational interventions. Publication bias is common in the social sciences, and there is no reason to believe that gifted education is any different. It is likely that every meta-analysis in the field (including the Rogers, 2007, study) is tainted by an unknown level of publication bias. The sooner gifted education scholars and practitioners admit this, the sooner they can remedy the problem of publication bias.

Second, researchers, peer reviewers, and editors should cease to think of rejecting the null hypothesis as the equivalent of a "successful" study and retaining the null hypothesis as a "failure." This mistaken belief leaders peer reviewers and editors to recommend rejection for manuscripts that retain the null hypothesis. It is unclear how common this belief is among gifted education scholars, but I have encountered peer reviewers in related fields (e.g., school psychology) who have this erroneous belief. The mistaken belief that rejecting the null hypothesis is always desirable also leads researchers to not submit their studies to journals, which leads to the infamous "file drawer problem" where an unknown number of studies are squirreled away in researchers' file drawers, forever hidden from public view (Rosenthal, 1979). The result is a distorted literature that contains a surplus of statistically significant results and positively biased estimates of effect sizes.

Third, researchers, reviewers, and editors should work under the assumption that every study is worth publishing in some venue. All pilot studies (Meehl, 1990), replications (Makel & Plucker, 2014, 2015), and local program evaluation studies—all of which have traditionally been underrepresented in the literature—should be disseminated to the widest audience possible. This is not an invitation to eliminate publication standards in *Gifted Child Quarterly* or any other journal. However, criteria that exasperate publication bias should not be considered in the publication decision. Some of these criteria include a study's novelty, the magnitude of its effect size, and whether the research questions are "interesting." Rather, every ethically and methodologically sound study warrants publication. If a study is still rejected by a journal (or was never worthy of publication, perhaps because of pervasive problems with internal validity), then it should still be disseminated publically, and authors should deposit their work in online archival sites, such as ERIC, researchgate.net, academia.edu, the Social Science Research

Network, and the newly launched SocArXiv and PsyArXiv. In this way other scholars can discover their work, and the problem of publication bias will be reduced.

Fourth, researchers need to increase the statistical power of their studies. The statistical power calculations in Tables 1 and 2 show that few studies meet the suggested .80 level of power (Cohen, 1962) when an effect size is of "medium" magnitude (i.e., d = .50). For example, for the between-subjects designs the average power for studies of academic achievement outcomes was .4523 for a medium effect size. This means that the odds that a study would reject the null hypothesis for a medium strength effect size were about as good as flipping a coin. Although this rate is comparable to other studies of statistical power (e.g., Cohen, 1962; Osborne, 2008; Sedlmeier & Gigerenzer, 1989), it still shows how poorly equipped most studies on academic enrichment are for detecting an effect of a program. Future researchers should try—whenever possible—to increase statistical power. This is not an easy recommendation to follow; the required sample sizes may be surprising to some. Assuming the use of a between-subjects design, the required sample size to detect a population effect size of d = .50 effect with .80 power is 63 individuals in each group. However, if the population effect size is d = .20, then the required effect sample size rises to 392 individuals in each group. Table 1 shows that in the enrichment literature only one study on academic achievement outcomes and three studies on other outcomes met the former threshold, and only one study (Olenchek, 1990) met the latter threshold. Although this may seem daunting, it is necessary because one large study provides more stable information—and is less susceptible to publication bias—than a collection of smaller studies (Schimmack, 2012).

Finally, gifted education should focus on performing meta-analyses with larger numbers of studies in order to make publication bias easier to detect. Rogers's meta-analyses of grouping

and acceleration interventions had a median of 15 eligible studies (mean = 19.0, SD = 14.9) eligible, which shows that Kim's (2016) meta-analysis is probably a typical meta-analysis for gifted education, with 13 effect sizes for academic achievement outcomes and 18 effect sizes for socioemotional outcomes. Thus, it is likely difficult to detect publication bias in most gifted education meta-analyses. Increasing the number of studies in a meta-analysis may be the hardest suggestion of all to implement because gifted education is a theoretically fractured field with few unifying ideas (Ambrose, Van Tassel-Baska, Coleman, & Cross, 2010), and research funding is scare compared to other branches of the social sciences. Nevertheless, I believe that an initiative from the National Association for Gifted Children (NAGC) encouraging researchers to target specific research questions and/or highly desired replications would make later meta-analyses stronger. Incentives, such as guaranteed publication in *Gifted Child Quarterly*, reserved presentation time at NAGC's annual conference, funding, and co-authorship on a later meta-analysis, could help to reign in the maverick spirit of gifted education and help the field concentrate on building up a deeper research base on vital issues. In addition to making publication bias easier to detect, implementing this would give the field stronger answers to its most pressing questions, instead of a shallower understanding of many topics.

Limitations

Some of the limitations of my analyses have become apparent already in this manuscript. One problem is that most of my analyses are reliant on statistical significance tests, which are inherently sensitive to sample size. With relatively few studies in the enrichment literature, some of the procedures (e.g., the chi-squared tests, or the correlation between effect size and sample size) had lower statistical power than may be needed to detect publication bias. This shows the

difficulty of detecting publication bias and estimating its size and distorting effect on the scientific literature.

Another shortcoming of this analyses is that the IC was developed by Schimmack (2012) to examine multi-study articles on the same topic—not meta-analyses that were a collection of studies from many different researchers. However, I do not believe this is a barrier to the IC's use in this context. The equations and statistical concepts that the IC is built upon do not know whether a collection of studies all originated from a single article or from a collection of articles from many authors. The principles of statistical power are the same regardless of the origin of the studies. Thus, there is no conceptual or statistical difficulty with using the IC to evaluate a meta-analysis. One more substantive problem with using the IC is that there are no widely agreed upon standards for an IC that is "too high." More work needs to be done to establish guidelines for the use and interpretation of this new statistic.

A more important issue to be aware of is that an excess of statistical significant findings may not be due solely to publication bias (Ioannidis & Trikalinos, 2007). For years, social scientists have been aware that the little decisions researchers make as they collect, analyze, and report data can influence their final *p*-values (Greenwald, 1975; Kerr, 1998; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). These questionable research practices, such as selectively reporting results, adding or dropping covariates until results are statistically significant, or stopping data collection when the desired results are obtained, can drive down *p*-values and increase IC values. One could make an argument that these practices are not the same as publication bias—which would indicate that my results are not evidence of publication bias at all, but rather that "something else" is going on with how the studies in the enrichment literature are conducted. Yet, this argument ignores *why* a researcher would engage in questionable

research practices: to get the "right" results, which usually entails obtaining low *p*-values. Why do researchers want low *p*-values? Because of the bias in the social sciences among researchers, reviewers, and journal editors against the null hypothesis (Greenwald, 1975). Thus, even when the *immediate cause* of a surplus of rejected null hypotheses is due to questionable research practices, the *ultimate cause* is still publication bias. For those who wish to make the distinction between the file drawer problem and questionable research practices occurring before a manuscript is submitted to a journal, a survey of gifted education researchers asking about completed, unpublished studies could help scholars understand how much of the surplus of rejected null hypotheses is due to either cause.

Conclusion

Publication bias is a well-known problem in the social science, and many researchers can probably explain that it distorts the scientific literature. I have written this manuscript to show readers a contemporary, real-life example of how publication bias may have distorted the research base on a topic relevant to their substantive interests. I hope that this manuscript moves beyond the abstract warnings of the dangers of publication bias and instead serves as a wakeup call to researchers about how truly serious the problem of publication bias can be—as it likely distorts their knowledge of the phenomena they study. As the work of Kim (2016) demonstrates, the most diligent efforts of a careful meta-analyst cannot correct for this distortion. Statistical power analysis provides circumstantial evidence for the existence of publication bias in the enrichment literature. As a result, the true impacts of enrichment programs for gifted children are unknown—and likely cannot be known unless publication bias can be reduced.

Although Kim's (2016) surprisingly strong results for her meta-analysis were the inciting incident in this discussion of publication bias, it is unlikely that only the enrichment literature is

susceptible to this problem. I believe that the enrichment research is representative of many other branches of the educational psychology literature. Therefore, I suggest that researchers take active steps to reduce publication bias in gifted education and related fields. This includes disseminating unpublished studies, eliminating publication criteria that contribute to publication bias, increasing the statistical power of research studies, and increasing the number of studies in their meta-analyses. Some of these changes will be slow or may contradict years of professional practice, but these changes are necessary if gifted education is going to have a strong empirical research base. If gifted education scholars make these changes now, then the field may gain a reputation for producing trustworthy, high quality empirical research. As a result, gifted education can go back to being "incredible" in the sense of "exceedingly great," because of a strong, credible research literature.

References

Aldrich, P. W., & Mills, C. J. (1989). A special program for highly able rural youth in grades five and six. *Gifted Child Quarterly, 33*, 11-14. doi:10.1177/001698628903300102

Ambrose, D., Van Tassel-Baska, J., Coleman, L. J., & Cross, T. L. (2010). Unified, insular, firmly policed, or fractured, porous, contested, gifted education? *Journal for the Education of the Gifted, 33*, 453-478. doi:10.1177/016235321003300402

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Callahan, C. M., Moon, T. R., & Oh, S. (2014). *National surveys of gifted programs: Executive summary*. Retrieved from http://www.nagc.org/sites/default/files/key%20reports/2014%20Survey%20of%20GT%20programs%20Exec%20Summ.pdf

Callahan, C. M., Moon, T. R., Oh, S., Azano, A. P., & Hailey, E. P. (2015). What works in gifted education: Documenting the effects of an integrated curricular/instructional model for gifted students. *American Educational Research Journal, 52*, 137-167. doi:10.3102/0002831214549448

Carter, K. R. (1986). A cognitive outcomes study to evaluate curriculum for the gifted. *Journal for the Education of the Gifted, 10*, 41-55. doi:10.1177/016235328601000104

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in dducation. *Educational Researcher, 45*, 283-292. doi:10.3102/0013189x16656615

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*, 145-153. doi:10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003. doi:10.1037/0003-066x.49.12.997

Coleman, J. M., & Fults, B. A. (1982). Self-concept and the gifted classroom: The role of social comparisons. *Gifted Child Quarterly, 26*, 116-120. doi:10.1177/001698628202600305

Cooper, H. (2010). *Research synthesis and meta-analysis* (4th ed.). Thousand Oaks, CA: Sage.

Coronado-Montoya, S., Levis, A. W., Kwakkenbos, L., Steele, R. J., Turner, E. H., & Thombs, B. D. (2016). Reporting of positive results in randomized controlled trials of mindfulness-based mental health interventions. *PLoS ONE, 11*(4), e0153220. doi:10.1371/journal.pone.0153220

Durlak, J. A. (2009). How to select, calculate and interpret effect sizes. *Journal of Pediatric Psychology, 34*, 917-928. doi:10.1093/jpepsy/jsp004

Feldhusen, J. F., Sayler, M. F., Nielsen, M. E., & Kolloff, P. B. (1990). Self-concepts of gifted children in enrichment programs. *Journal for the Education of the Gifted, 13*, 380-387. doi:10.1177/016235329001300407

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53*, 25-44. doi:10.1016/j.jsp.2014.10.002

Gallagher, J. J. (2000). Unthinkable thoughts: Education of gifted students. *Gifted Child Quarterly, 44*, 5-12. doi:10.1177/001698620004400102

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology, 49*, 1886-1897. doi:10.1037/a0031412

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research, 37*, 3-30. doi:10.1177/0049124108318973

Ginsel, B., Aggarwal, A., Xuan, W., & Harris, I. (2015). The distribution of probability values in medical abstracts: An observational study. *BMC Research Notes, 8*, 721-725. doi:10.1186/s13104-015-1691-x

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8. doi:10.3102/0013189X005010003

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20. doi:10.1037/h0076157

Gross, M. U. M. (2004). *Exceptionally gifted children* (2nd ed.). New York, NY: Routledge.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*, 359-369. doi:10.1037/0033-2909.88.2.359

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power. *The American Statistician, 55*, 19-24. doi:10.1198/000313001300339897

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*, 245-253. doi:10.1177/1740774507079441

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196-217. doi:10.1207/s15327957pspr0203_4

Kim, M. (2016). A meta-analysis of the effects of enrichment programs on gifted students. *Gifted Child Quarterly, 60,* 102-116. doi:10.1177/0016986216630607

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Kolloff, M. B. (1983). *The effects of an enrichment program on the self-concepts and creative thinking abilities of gifted and creative elementary students* (Unpublished doctoral dissertation). Purdue University, West Lafayette, IN.

Kulik, J. A., & Kulik, C.-L. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly, 36,* 73-77. doi:10.1177/001698629203600204

Laws, K. R. (2013). Negativland - a home for all findings in psychology. *BMC Psychology, 1*(2), 1-8. doi:10.1186/2050-7283-1-2

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lubinski, D., Benbow, C. P., & Kell, H. J. (2014). Life paths and accomplishments of mathematically precocious males and females four decades later. *Psychological Science, 25,* 2217-2232. doi:10.1177/0956797614551371

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304-316. doi:10.3102/0013189x14545513

Makel, M. C., & Plucker, J. A. (2015). An introduction to replication research in gifted education: Shiny and new is not the same as useful. *Gifted Child Quarterly, 59,* 157-164. doi:10.1177/0016986215578747

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology, 65*, 2271-2279. doi:10.1080/17470218.2012.711335

Matthews, M. S. (2008). Talent Search programs. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education* (pp. 641-654). Waco, TX: Prufrock Press.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163. doi:10.1037/1082-989x.9.2.147

McCoach, D. B., Gubbins, E. J., Foreman, J., Rubenstein, L. D., & Rambo-Hernandez, K. E. (2014). Evaluating the efficacy of using predifferentiated and enriched mathematics curricula for grade 3 students: A multisite cluster-randomized trial. *Gifted Child Quarterly, 58*, 272-286. doi:10.1177/0016986214547631

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244. doi:10.2466/pr0.1990.66.1.195

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Quantitative Psychology and Measurement, 3*(102), 1-13. doi:10.3389/fpsyg.2012.00102

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology, 19*, 172-182. doi:10.1037/gpr0000034

Olenchak, F. R. (1990). School change through gifted education: Effects on elementary students' attitudes toward learning. *Journal for the Education of the Gifted, 14*, 66-78. doi:10.1177/016235329001400108

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*, 151-160. doi:10.1080/01443410701491718

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*, 1-12. doi:10.1037/a0018326

Rogers, K. B. (2007). Lessons learned about educating the gifted and talented: A synthesis of the research on educational practice. *Gifted Child Quarterly, 51*, 382-396. doi:10.1177/0016986207306324

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

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118-137. doi:10.1016/j.intell.2015.09.002

Ruf, D. L. (2005). *Losing our minds: Gifted children left behind*. Scottsdale, AZ: Great Potential Press.

Ruthsatz, J., Ruthsatz, K., & Ruthsatz Stephens, K. (2014). Putting practice into perspective: Child prodigies as evidence of innate talent. *Intelligence, 45*, 60-65. doi:10.1016/j.intell.2013.08.003

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551-566. doi:10.1037/a0029487

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540. doi:10.1037/0021-9010.62.5.529

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316. doi:10.1037/0033-2909.105.2.309

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

Slade, M. K., & Warne, R. T. (2016). A meta-analysis of the effectiveness of trauma-focused cognitive-behavioral therpay and play therapy for child victims of abuse. *Journal of Young Investigators, 30*(6), 36-43.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760. doi:10.1037/0003-066x.32.9.752

Stanley, J. C. (1976). The case for extreme educational acceleration of intellectually brilliant youths. *Gifted Child Quarterly, 20*, 66-75. doi:10.1177/001698627602000120

Stanley, J. C. (1977). Rationale of the Study of Mathematically Precocious Youth (SMPY) during its first five years of promoting educational acceleration. In J. C. Stanley, W. C. George & C. H. Solano (Eds.), *The gifted and the creative: A fifty-year perspective* (pp. 75-112). Baltimore, MD: The Johns Hopkins University Press.

Steenbergen-Hu, S., & Olszewski-Kubilius, P. (2016). How to conduct a good meta-analysis in gifted education. *Gifted Child Quarterly, 60,* 134-154. doi:10.1177/0016986216629545

te Nijenhuis, J., Willigers, D., Dragt, J., & van der Flier, H. (2016). The effects of language bias and cultural bias estimated using the method of correlated vectors on a large database of IQ comparisons between native Dutch and ethnic minority immigrants from non-Western countries. *Intelligence, 54,* 117-135. doi:10.1016/j.intell.2015.12.003

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434-438. doi:10.1002/j.1556-6676.1992.tb01631.x

Vadillo, M. A., Hardwicke, T. E., & Shanks, D. R. (2016). Selection bias, vote counting, and money-priming effects: A comment on Rohrer, Pashler, and Harris (2015) and Vohs (2015). *Journal of Experimental Psychology: General, 145,* 655-663. doi:10.1037/xge0000157

VanTassel-Baska, J. (1987). The ineffectiveness of the pull-out program model in gifted education: A minority perspective. *Journal for the Education of the Gifted, 10,* 255-264.

Vaughn, V. L., Feldhusen, J. F., & Asher, J. W. (1991). Meta-analyses and review of research on pull-out programs in gifted education. *Gifted Child Quarterly, 35,* 92-98. doi:10.1177/001698629103500208

Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2010). Accomplishment in science, technology, engineering, and mathematics (STEM) and its relation to STEM educational dose: A 25-year longitudinal study. *Journal of Educational Psychology, 102,* 860-871. doi:10.1037/a0019454

Wai, J. (2015). Long-term effects of educational acceleration. In S. G. Assouline, N. Colangelo, J. VanTassel-Baska & A. Lupkowski-Shoplik (Eds.), *A nation empowered: Evidence trumps the excuses holding back America's brightest students* (Vol. 2, pp. 73-83). Iowa City, IA: Belin-Blank Center.

Warne, R. T. (2011). A reliability generalization of the Overexcitability Questionnaire–Two. *Journal of Advanced Academics, 22*, 671-692. doi:10.1177/1932202x11424881

Warne, R. T. (2016). Five reasons to put the *g* back into giftedness: An argument for applying the Cattell–Horn–Carroll theory of intelligence to gifted education research and practice. *Gifted Child Quarterly, 60*, 3-15. doi:10.1177/0016986215605360

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist, 4*, 140-143. doi:10.2307/27701483

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30*, 141-167. doi:10.3102/10769986030002141

Zimmerman, D. W. (1997). Teacher's corner: A note on interpretation of the paired-samples *t* test. *Journal of Educational and Behavioral Statistics, 22*, 349-360. doi:10.3102/10769986022003349

Footnotes

1. Indeed, if Kim had not been so clear in her explanation of her decisions and procedures, then it would have been impossible for me to investigate the cause of her high effect sizes—much less write an entire article about it.
2. Other study characteristics can strengthen observed effect sizes slightly, such as higher reliability of variables and stronger treatment fidelity (Nimon, Zientek, & Henson, 2012), but these adjustments have usually have relatively minor effects on statistical power.
3. In fact, if one believes that the population effect size is $d = 0$, then IC values based on this estimate will be equal to the *post hoc* statistical power. This would indicate an estimated level of publication bias so rampant that *every* rejected null hypothesis would be a Type I error.
4. If forced to take a guess, I would have estimated that the effect sizes for the academic benefits of enrichment programs would be between $d = .20$ and $d = .50$. For socioemotional benefits, I would have estimated that the effect sizes would be between $d = 0$ and $d = .30$.

Table 1
Effect Sizes and Statistical Power of Studies on Academic Impacts of Enrichment Programs

Study	*n*	Design	*d*	$p < .05$?	*Post hoc* power	Power for mean ES in meta-analysis ($d = 0.96$ or 0.65)[a]	Power for medium effect size ($d = 0.50$)	Power for Rogers's (2007) median effect size ($d = 0.45$)
Aldrich & Mills (1989)	77	Between groups	0.58	Yes	0.5933	> 0.9999	0.4868	0.4108
Aljughaiman (2011)	88	Within groups	0.73	Yes	> 0.9999	> 0.9999	0.9963	0.9866
Aljughaiman & Ayoub (2012)	42	Between groups	2.75	Yes	> 0.9999	0.8427	0.3675	0.3086
Delaney (1978)	60	Between groups	0.77	Yes	0.7342	0.5883	0.3898	0.3274
Farleigh-Lohrfink et al. (2013)	66	Between groups	2.05	Yes	> 0.9999	0.9292	0.5235	0.4433
Gubbels et al. (2014)	66	Between groups	0.38	No	0.3249	0.9738	0.5291	0.4483
Lee et al. (2010)	45	Within groups	0.69	Yes	0.9926	> 0.9999	0.9069	0.8398
Lynch & Mills (1990)	64	Between groups	2.99	Yes	> 0.9999	0.9701	0.5171	0.4376
Mahmoud (2014)	30	Within groups	1.00	Yes	0.9993	> 0.9999	0.7548	0.6644
McCoach et al. (2014)	301	Between groups	0.41	Yes	0.9424	> 0.9999	0.9906	0.9726
Miller & Gentry (2010)	54	Between groups	0.12	No	0.0704	0.9144	0.4523	0.3454
Newman (2005)	54	Between groups	0.95	Yes	0.9278	0.9418	0.4523	0.3808
Sastre-Riba (2013)	49	Within groups	1.28	No	> 0.9999	> 0.9999	0.9293	0.8703

Shi et al. (2013)	48	Between groups	0.80	Yes	0.7718	0.7718	0.4111	0.3454
Wilson (1959)	25	Between groups	-0.10	No	0.0575	0.3695	0.2401	0.2034
Median power (all effect sizes):					0.9424	0.9701	0.5171	0.4376
Incredibility Index:					-0.2091	-0.2368	0.2162	0.2957
Median power (effect sizes from between groups designs):					0.7718	0.9292	0.4523	0.3808
Incredibility Index:					0.0464	-0.1110	0.3659	0.4374

[a]Mean effect size was d = .96 for effect sizes that appeared in Kim's (2016) meta-analysis. For the two studies that only appeared in Vaughn et al.'s (1991) meta-analysis (i.e., Delaney, 1978; Wilson, 1959), d = .65.

Table 2
Effect Sizes and Statistical Power of Studies on Academic Impacts of Enrichment Programs

Study	*n*	Design	*d*	$p < .05$?	*Post hoc* power	Power for mean ES in meta-analysis (*d* varies)	Power for medium effect size (*d* = .50)	Power for Rogers's (2007) median effect size (*d* = .15)
Alujughaiman (2011)	88	Within groups	2.61	Yes	> 0.9999	0.9992[a]	0.9963	0.2849
Beckwith (1982)	43	Between groups	0.44	No	0.3037	0.3037[b]	0.3754	0.3152
Carter (1986)	48	Between groups	0.24	No	0.1328	0.3328[b]	0.4111	0.3454
Cho & Lee (2006)	76	Within groups	0.49	Yes	0.9850	0.9972[a]	0.9904	0.2518
Cohen et al. (1994)	202	Between groups	0.50	Yes	0.8788	0.9308[a]	0.8788	0.1559
Coleman & Fults (1982)	134	Between groups	-0.36	Yes[c]	0.5489	0.0979[d]	0.8242	0.1401
Cunningham & Rinn (2007)	140	Within groups	0.50	Yes	> 0.9999	> 0.9999[a]	> 0.9999	0.4218
Dai et al. (2012)	291	Within groups	0.02	No	0.0526	> 0.9999[a]	> 0.9999	0.7225
Feldhusen et al. (1990), elementary	38	Between groups	0.57	Yes	0.3967	0.9740[a]	0.3193	0.0735
Feldhusen et al. (1990), middle	22	Between groups	0.34	No	0.1031	.2104[a]	0.1820	0.0617
Fraleigh-Lohrfink et al. (2013)	66	Between groups	0.67	Yes	0.7593	.6028[a]	0.5235	0.0933
Gubbels et al. (2014)	66	Between groups	0.68	No	0.7773	.6087[a]	0.5291	0.0939
Hay et al. (2000)	20	Between groups	13.42	Yes	> 0.9999	.2342[a]	0.2018	0.0634

Kolloff (1983), elementary	392	Between groups	0.13	No	0.2519	—[e]	0.9986	0.3186
Kolloff & Moore (1989), middle	439	Within groups	0.81	Yes	> 0.9999	> 0.9999[a]	> 0.9999	0.8802
Kolloff & Moore (1989), high school	69	Within groups	0.79	Yes	> 0.9999	0.9945[a]	0.9836	0.2323
Olenchak (1990)	1935	Between groups	0.01	No	0.0559	> 0.9999[a]	> 0.9999	0.9095
Olenchak (1995)	108	Within groups	0.82	Yes	> 0.9999	> 0.9999[a]	0.9993	0.3391
Phillips et al. (2002)	32	Within groups	0.19	No	0.1653	0.8546[a]	0.7830	0.1277
Rinn (2006)	140	Within groups	0.30	Yes	0.9412	> 0.9999[a]	> 0.9999	0.4218
Stake & Mares (2001)	330	Within groups	0.04	No	0.1083	> 0.9999[a]	> 0.9999	0.7755
Stake & Mares (2005)	88	Within groups	0.03	No	0.0463	0.9992[a]	0.9963	0.2849
Median power (all effect sizes):					0.6541	0.9945	0.9870	0.2849
Incredibility Index:					-0.1541	-0.4945	-0.4870	0.2151
Median power (effect sizes from between groups designs):					0.3967	0.3534	0.5235	0.1401
Incredibility Index:					-0.0331	0.0102	-0.1599	0.2235

[a]Mean effect size was $d = 0.55$ for these socioemotional effect sizes in Kim's (2016) meta-analysis.
[b]Mean effect size was $d = 0.44$ for these critical thinking effect sizes in Vaughn et al.'s (1991) meta-analysis.
[c]This result was statistically significant, but not in the expected direction. Therefore, it is not counted as demonstrating publication bias.
[d]Mean effect size was $d = 0.11$ for the socioemotional effect sizes in Vaughn et al.'s (1991) meta-analysis.
[e]These effect sizes combine effect sizes from multiple areas (e.g., creativity and socioemotional).