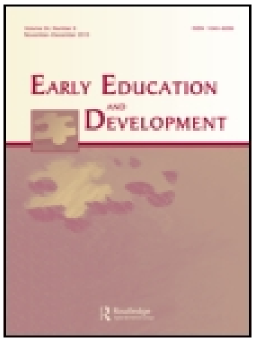

Early Education and Development

ISSN: 1040-9289 (Print) 1556-6935 (Online) Journal homepage: http://www.tandfonline.com/loi/heed20

Understanding the Program Effectiveness of Early Mathematics Interventions for Prekindergarten and Kindergarten Environments: A Meta-Analytic Review

Aubrey H. Wang, Janine M. Firmender, Joshua R. Power & James P. Byrnes

To cite this article: Aubrey H. Wang, Janine M. Firmender, Joshua R. Power & James P. Byrnes (2016) Understanding the Program Effectiveness of Early Mathematics Interventions for Prekindergarten and Kindergarten Environments: A Meta-Analytic Review, Early Education and Development, 27:5, 692-713, DOI: 10.1080/10409289.2016.1116343

To link to this article: http://dx.doi.org/10.1080/10409289.2016.1116343

EARLY EDUCATION AND DEVELOPMENT
2016, VOL. 27, NO. 5, 692–713
http://dx.doi.org/10.1080/10409289.2016.1116343

Taylor & Francis
Taylor & Francis Group

Understanding the Program Effectiveness of Early Mathematics Interventions for Prekindergarten and Kindergarten Environments: A Meta-Analytic Review

Aubrey H. Wang[a], Janine M. Firmender[b], Joshua R. Power[a], and James P. Byrnes[c]

[a]Educational Leadership, Saint Joseph's University; [b]Teacher Education, Saint Joseph's University; [c]Educational Psychology, Temple University

ABSTRACT

Research Findings: The early childhood years are critical in developing early mathematics skills, but the opportunities one has to learn mathematics tend to be limited, preventing the development of significant mathematics learning. By conducting a meta-analysis of 29 experimental and quasi-experimental studies that have been published since 2000, this study extends beyond prior evaluations of early mathematics programs for prekindergarten and kindergarten environments by quantifying program effectiveness in terms of effect sizes and examining 6 aspects of these programs. We found an overall moderate to large effect size. There was a tendency for programs to produce larger effects when they (a) targeted a single content strand, (b) presented content 120 to 150 min per week, (c) designed programs for the prekindergarten environment, (d) presented content individually to children, and (e) used researcher-made mathematics assessments. *Practice or Policy*: Early mathematics programs can be designed to be both developmentally appropriate and highly effective. The goal of this meta-analysis was to reveal these programs to early development and education professionals so that they understand some of the factors that might explain why some produced stronger effects than others.

A renewed emphasis has been placed on mathematics education and early childhood education in the United States as results from both national and international assessments indicate that students in the United States have not performed as well as students from other countries in mathematics (Mullis, Martin, Foy, & Arora, 2012). At the international level, U.S. fourth graders rank below their same-age peers from eight countries, and U.S. eighth graders rank below their peers in 11 countries (Mullis et al., 2012). At the national level, the majority of fourth graders (60%) and eighth graders (57%) failed to meet standards of proficiency in mathematics (National Center for Education Statistics, 2011).

Although it is common to attribute these disconcerting achievement outcomes to deficiencies in the instructional practices of elementary and middle school teachers in the United States, a number of recent studies suggest that the problem may actually originate in the early childhood period. Even though the first 5 years of life are critical in developing children's early mathematics skills (Byrnes & Wasik, 2009; Duncan, Brooks-Gunn, & Klebanov, 1994; National Institute of Child Health and Human Development Early Child Care Research Network, 2006; Wang, Shen, & Byrnes, 2013), very little mathematics is normally presented during the prekindergarten (pre-K) and kindergarten years (Early et al., 2005; Hausken & Rathbun, 2004). In a study of 11 state-funded pre-K programs, children were observed taking part in math-related activities only 8% of the time (Early et al., 2005).

CONTACT Aubrey H. Wang ✉ awang@sju.edu Educational Leadership, Saint Joseph's University, Merion Hall 266, 5600 City Avenue, Philadelphia, PA 19131.

The time spent on learning mathematics in kindergarten does seem to increase; however, the time varies between full-day and half-day programs. For example, it has been reported that in full-day kindergarten teachers provide about 3.6 hr per week of mathematics instruction, whereas half-day kindergarten teachers provide about 2.4 hr per week (Hausken & Rathbun, 2004). This means that on average kindergarten students are engaged in mathematics instruction for approximately 3.1 hr a week. In comparison, kindergarten students are engaged in reading instruction for about 5.2 hr per week (Hausken & Rathbun, 2004). When time is not devoted for students to engage in mathematics tasks, talk, and/or play, significant mathematics learning is not likely to occur (Wang, 2010). In 2009, the National Research Council's Committee on Early Childhood Mathematics concluded that "most early childhood programs spend little focused time on mathematics, and most of it is low in instructional quality" (p. 339).

One obvious policy implication of such findings is that more time should be devoted to mathematics during the early childhood years. However, there are many ways to infuse more mathematical content into early childhood environments. From what has been learned from cognitive and developmental science, some approaches would be expected to be more effective and developmentally appropriate than others (Byrnes, 2008; National Association for the Education of Young Children [NAEYC], 2009). In addition, any decision to alter instructional practices should be based on sound empirical evidence using guidelines such as those established by the Institute of Education Sciences (Frye et al., 2013).

Fortunately, a number of early childhood math programs have been created over the past 15 years that to a large extent are grounded in developmental science and also developmentally appropriate. Moreover, many of these programs have been evaluated using rigorous methodologies such as randomized controlled experiments. At the request of the Institute of Education Sciences, Frye et al. (2013) summarized 29 studies that evaluated early childhood mathematics programs and made five recommendations for practice that derived from expert opinion about the characteristics of high-quality programs. Their goal was to report the level of support that these studies collectively provided for each recommendation using What Works Clearinghouse criteria. Frye et al. reported that there was moderate evidence to support the first recommendation that number and operations should be taught according to a developmental progression. In contrast, they found minimal evidence in these studies to support the remaining four recommendations: (a) providing instruction in geometry, patterns, measurement, and data analysis using a developmental progression (Recommendation 2); (b) implementing a progress-monitoring approach that establishes a child's level of knowledge, tailors instruction to the child's individual needs and developmental level, and monitors progress to facilitate the child's building connections between new math knowledge and what the child already knows (Recommendation 3); (c) teaching children to view and describe their world mathematically (Recommendation 4); and (d) providing math instruction daily as well as taking advantage of opportunities to integrate math into other classroom activities, including games and instruction in other content areas (Recommendation 5). Thus, even though the studies reviewed were well designed, the programs themselves were not informed in advance by four of the five recommendations and the studies could not therefore provide evidence in support of the recommendations.

Although the report of Frye et al. (2013) is certainly useful for identifying the aforementioned studies and showing the degree to which the design of the evaluated programs is consistent with expert opinion, in the present study we considered whether there were additional ways to distill useful information from these and other studies that recently evaluated mathematical interventions for early childhood settings. Our goal was to provide effect size information related to aspects of these programs (e.g., whether more comprehensive programs produced larger effects than less comprehensive programs) in order that early childhood educators could make more informed decisions about whether it is worth the effort to implement particular programs and which features of programs produced the largest effects. In addition, we expected that the effect size findings for program aspects would also be informative theoretically as well. As we discuss later, some findings would confirm theoretical expectations, and some would be inconsistent with theoretical expectations.

Although we initially considered a number of possible features of programs (e.g., the degree to which a program was consistent with NAEYC standards), there were often not enough programs in particular categories to conduct meaningful comparisons. In the end, we examined the following program dimensions that had relevance to both theory and practice: (a) whether programs were supplemental activities added to an ongoing curriculum versus comprehensive, stand-alone curricula; (b) whether a single National Council of Teachers of Mathematics (NCTM) mathematics content strand (NCTM, 2000) was targeted (e.g., number and operations) or multiple content strands were present (e.g., number and operations, algebra, geometry, data analysis and probability, measurement); (c) the amount of time devoted to math activities per week; (d) whether programs targeted children in pre-K programs or kindergarten programs; (e) whether content was presented to children individually, in small groups, or in a whole-class format; and (f) whether the outcome measure was a researcher-made assessment or a standardized math test. In what follows, we briefly discuss the practical and theoretical implications of these six comparisons.

Six meta-analytic comparisons and their implications

With regard to the first program feature evaluated in this meta-analysis, given the fact that so little time is normally devoted to mathematics in many early childhood environments (as argued previously), it stands to reason that early childhood educators would be more inclined to adopt a program that was a supplement to their ongoing curriculum than adopt a more comprehensive program with many features and facets (Siegler & Ramani, 2008). The latter is likely to be more costly as well and require more professional development and training. If the present meta-analysis were to reveal comparable effect sizes for the supplemental activity and more comprehensive programs, early childhood educators might reasonably opt for the former. However, such a finding would also be somewhat surprising on a theoretical level as well and would elicit a variety of post hoc explanations that could be examined in follow-up studies. With young children, there is likely an optimal level of exposure in which topics recur with sufficient frequency and intensity so as to engender learning but not so much as to be taxing (Byrnes, 2008; NAEYC, 2009). Too much exposure could yield diminishing returns, especially if not implemented as part of a child-centered, playful program (Fisher, Hirsh-Pasek, Newcombe, & Golinkoff, 2013).

Similar arguments can be made for the second kind of comparison considered in the present meta-analysis: whether programs that target multiple content strands outlined by the NCTM (e.g., number and operations, geometry, and measurement) yield stronger effects than programs that target a single content strand (number and operations only). However, in this case it would be necessary to add the additional requirement that the outcome measures used to evaluate the two kinds of programs would include items reflecting multiple content strands. If the single-strand programs used assessments that only required knowledge of that one strand, whereas the multiple-strand programs used assessments that required knowledge of multiple strands, comparable effect sizes would not mean that the two kinds of programs had comparable effects. It would, however, be important and surprising to see transfer effects from programs that target a single strand to improved performance on other strands. Such transfer effects could occur if a program targeted a foundational mathematical ability that has a clear neuroscientific basis, such as an ordinal or linear representation of numbers (Ramani & Siegler, 2008). And as argued by the NCTM in its 2006 document *Curriculum Focal Points for Prekindergarten Through Grade 8 Mathematics*, instruction should focus on specific content strands at each grade level from pre-K to Grade 12. By focusing instruction “on a small number of key areas of emphasis, students gain extended experience with core concepts and skills. Such experience can facilitate deep understanding, mathematical fluency, and an ability to generalize" (p. 5).

With respect to the third program feature evaluated in this meta-analysis (i.e., the time devoted to mathematics—e.g., 30 min per week vs. 60 min per week), the issue again may be more complicated than it initially appears. At first blush, it may seem reasonable to assume that a program that exposes children to more content or more practice would yield stronger effects than a program that exposes children to less content or less practice. However, there is a growing precedent in the early childhood language literature showing that more exposure to vocabulary words by teachers or mothers does not seem to lead to faster growth in children's vocabulary in a straightforward manner (e.g., Rowe, 2012). Language researchers have been arguing and finding that it is not the quantity but the quality that matters (Rowe, 2012; Wasik & Hindman, in press). Thus, should our meta-analysis reveal that programs that devote less time to math yield comparable effects to those that devote more time, such a finding should generate interest in conducting follow-up studies that contrast the quantity versus quality of exposure.

The fourth issue examined in the present analysis pertains to whether programs were designed to be implemented in pre-K environments versus kindergarten environments. Once again, there are several contrasting hypotheses that could be proposed in advance of examining the results via meta-analysis. On the one hand, pre-K environments are designed with the more limited cognitive, behavioral, and attentional capacities of 3- and 4-year-olds in mind (Byrnes, 2008; Van De Walle, Lovin, Karp, & Bay-Williams, 2013). There is also significant development in the math abilities of children between the pre-K and kindergarten years (Mix, Huttenlocher, & Levine, 2002). Thus, one could argue that programs designed for children in kindergarten classrooms might be more effective than programs designed for children in pre-K environments because older children might be more cognitively ready for this instruction and the structure of kindergarten classrooms might allow for a more seamless assimilation of the program. On the other hand, it is sometimes easier to produce stronger growth in children who know less, so it is conceivable that programs for younger children may yield stronger effects. Once again, it will be useful and interesting to examine the findings of the meta-analysis to see which of these two proposals seems to be correct.

With respect to the fifth factor examined, the size of the group that is presented the content, there are several reasons to expect that (a) programs that are delivered to children individually would generate larger effects than programs that deliver content to children in small groups and (b) programs that deliver content in small groups would generate larger effects than programs delivered to an entire class. The first reason is that individual or small-group formats should work best because of the limited cognitive, behavioral, and attentional capacities of young children (Byrnes, 2008; Van De Walle et al., 2013). Second, working with students in a small-group setting allows a teacher to better assess students' mathematical understanding (Legnard & Austin, 2012). Third, meta-analyses of class size effects support these expectations (Hedges & Stock, 1983). However, the latter findings were primarily based on business-as-usual classrooms rather than randomized controlled interventions.

Finally, it would normally be expected that effect sizes should be larger when student mathematical knowledge is assessed with outcome measures that were designed by the researchers themselves than with standardized math assessments. Researcher-made assessments typically include items that closely match the content and skills fostered in particular interventions. In contrast, standardized assessments may emphasize content and skills that are not always included in an intervention. Should it be found that effect sizes are comparable regardless of the assessment, such a finding would suggest that some generalization has taken place.

In sum, the primary purpose of the present meta-analysis was to extend beyond prior evaluations of early childhood programs (e.g., Frye et al., 2013) by quantifying program effectiveness in terms of effect sizes and examining six aspects of programs that have both practical and theoretical value. Our research questions were the following:

(1) Across all programs, what is the average effect size of programs designed to improve the math skills of young children?
(2) Which aspects of programs seem to be associated with larger effects? The contrasts include (a) comprehensive curricula versus supplemental activities, (b) targeting multiple content strands versus targeting a single content strand, (c) programs designed for pre-K environments versus programs designed for kindergartners, and (d) content presented individually to children versus in small groups versus in whole groups.
(3) Do programs that devote larger amounts of time to math activities produce larger effect sizes than programs that devote less time?
(4) Are effect sizes larger when the outcome measure was designed by the researchers implementing the intervention than when the outcome measure is a standardized math assessment?

Method

Data sources

We used a three-phase search process to identify studies to be included in the meta-analysis. During Phase 1 we conducted a systematic search of academic databases, organization websites, and major journals. Specifically, we conducted keyword searches of abstracts in three academic databases—Academic Search Premier (via EBSCO), JSTOR, and ProQuest—using the following search terms: *early childhood math* OR *preschool math* OR *kindergarten math* AND *effective math learning opportunities*. We also searched for publications relating to early childhood mathematics on the websites of three well-known organizations that conduct work on early childhood education and/or provide analyses of relevant work: Child Care and Early Education Research Connections, the National Institute for Early Education Research, and What Works Clearinghouse. Lastly, we conducted hand searches of four prominent education research journals: *American Education Research Journal, Early Childhood Education Journal, Early Childhood Research Quarterly*, and *Journal for Research in Mathematics Education*. All search strategies were limited to the years 2000 to the present and yielded 1,433 articles and reports.

Studies were then identified for initial inclusion if a review of their abstracts revealed that they met the following criteria: (a) The document reported on original research; (b) the research was about school-based early mathematics learning opportunities (e.g., programs, curricula, supplemental activities) for pre-K or kindergarten students; (c) the opportunities being studied were linked to some student outcome (e.g., test scores, skills, interest); and (d) the research was published in an academic journal, in a book, in a conference proceeding, or by an organization. These criteria represented a desire to include relevant research appearing in a range of publication channels. To ensure that each of us applied the criteria reliably, we conducted an interrater test that resulted in 80% agreement. The average kappa measure of agreement value was 0.5 ($p < .001$), or moderate agreement (Peat, 2001), and was considered acceptable. We reviewed the discrepant decisions, and more precise rules were created. These rules were documented and used to guide the inclusion decisions for all studies. We identified 80 studies for the initial sample at the end of Phase 1.

During Phase 2, each of us reviewed the entirety of each study closely and provided a summary of the study's definition of the mathematics program, research design, assessment instrument, and results. This process revealed that on closer review, 25 studies from the initial sample did not meet the inclusion criteria and were eliminated. For the remaining 55 studies, we conducted a citation analysis to make sure that additional relevant studies were not overlooked. Studies that were cited three or more times in the references of the 55 included studies, had been published since 2000, and met the inclusion criteria for this study were identified. Of the more than 3,000 citations analyzed, four studies met these criteria and were subsequently included in the sample. In addition, we identified two more studies for inclusion based on their review of the most recent issues of relevant academic journals. At the end of Phase 2, 61 studies were included in the sample.

During Phase 3, only studies that utilized randomized controlled trials (randomly assigned participants) or quasi-experimental designs (not randomly assigned participants) in which (a) treatment groups received a particular math program and control groups received business-as-usual math instruction and (b) an (experimenter-made or standardized) outcome measure was administered that assessed math skills were included for further analysis. At the end of this phase, a total of 24 studies remained. Given the small number of studies ultimately identified for inclusion in the current meta-analysis ($n = 24$), we decided to widen the search to locate more studies that contained a treatment group and control group for early math programs by not limiting our search to only studies conducted after 1999. We found five additional studies that met our criteria, but not surprisingly, all of the new studies were conducted after 2000. Table 1 provides a summary of key characteristics of the 29 studies included.

Analysis for meta-analytic review

We used Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, n.d.) to conduct our meta-analysis. Because all 29 studies used either an experimental ($n = 22$) or

Table 1. Summary of characteristics of studies included in this meta-analysis ($n = 29$).

Study characteristic	*n*	%
Independent variable	10	34
Curriculum	19	66
Supplemental mathematics-related activity		
Dependent variable[a]	16	50
General math achievement	13	41
Numeracy	3	9
Other		
Sample size	14	48
<100	7	24
100–200	4	14
201–500	4	14
>500		
Study design	7	24
Quasi-experimental	22	76
Randomized controlled		
Reliability of the dependent variable[b]	1	3
.60–.80	20	69
.80–1.0	8	28
Not reported		
Sample race/ethnicity		
>70% majority and mixed	9	31
>70% minority	14	48
International	5	17
Not reported	1	3
Sample socioeconomic status		
Low	16	55
Mixed	12	41
Not reported	1	3
Sample location		
Urban	14	48
Mixed	4	14
Not reported	11	38

[a] Results total more than 100% because some studies used multiple dependent variables measuring mathematics achievement.
[b] Results total more than 100% because some studies reported reliability statistics for multiple dependent variables measuring mathematics achievement.

quasi-experimental design (n = 7), all of the effect sizes were calculated from each study based on a comparison of an experimental group and a control group. An effect size is the standardized measure of association between two variables that is used by researchers in meta-analyses to compare effects of similar treatment across various studies (Borenstein, Hedges, Higgins, & Rothstein, 2009). For the 10 studies that reported means and standard deviations for independent groups (Aunio, Hautamaki, & Van Luit, 2005; Casey et al., 2008; Curtis, Okamoto, & Weckbacher, 2009; Dyson, Jordan, & Glutting, 2013; Kidd, Pasnak, Gadzichowski, Ferral-Like, & Gallington, 2008; Monahan, 2007; Pasnak, 2006; Sarama, Clements, Starkey, Klein, & Wakeley, 2008; Starkey, Klein, & Wakeley, 2004; Young-Loveridge, 2004), the mean difference between the experimental and control groups was the numerator and the pooled standard deviation was the denominator (Borenstein et al., 2009). For the 14 studies that reported pre/post mean and standard deviation (Arnold, Fisher, Doctoroff, & Dobbs, 2002; Chard et al., 2008; Clarke et al., 2011; Clements & Sarama, 2007, 2008; Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Fuchs, Fuchs, & Karns, 2001; Jordan, Glutting, Dyson, Hassinger-Das, & Irwin, 2012; Klein, Starkey, Clements, Sarama, & Iyer, 2008; Papic, Mulligan, & Mitchelmore, 2011; Sarama, Lange, Clements, & Wolfe, 2012; Sood, 2009; Sophian, 2004; Tarim, 2009), the mean pre/post change in the treatment group minus the mean pre/post change in the control group was the numerator and the pooled pretest standard deviation was the denominator (Morris, 2008). For the studies that reported F-test results (Baroody, Eiland, & Thompson, 2009; Pagani, Jalbert, & Girard, 2006; Ramani & Siegler, 2008, 2011; Siegler & Ramani, 2009), the standardized mean difference (d) and 95% confidence intervals were calculated using Wilson's (n.d.) practical meta-analysis effect size calculator available from the Campbell Collaboration website. This included calculating effect sizes for F-tests for two groups of unequal sample size and F-tests with three or more groups.

In all, 31 effect sizes were calculated from the 29 studies that were included in the meta-analysis. Consistent with the meta-analysis of Manz, Hughes, Barnabas, Bracaliello, and Ginsburg-Block (2010), we calculated an effect size for each individual intervention group in the study rather than calculating a mean effect size to represent the multiple early math intervention groups. Although this procedure resulted in effect sizes that were not independent, it was beneficial for a detailed examination of the effectiveness of various early mathematics intervention components.

Random-effects model

Meta-analyses' effect sizes generally are calculated based on either the fixed-effect model or the random-effects model (Borenstein et al., 2009; Cooper & Hedges, 1994; Stanley & Doucouliagos, 2015). A fixed-effect model assumes that the effect sizes calculated are adequate for estimating the one true effect size that can be generalized to other studies (Borenstein et al., 2009; Lipsey & Wilson, 2001) and that all differences in the observed effect sizes are due to sampling error. A paradigmatic example of studies that match this assumption is a multisite medical study examining the effect of a particular drug in which all aspects of the protocol (e.g., dosage, time of administration, method of administration) are intentionally kept the same. By contrast, under a random-effects model, each study is assumed to have its own true effect size to reflect population differences in participant demographics, contextual factors, variable program aspects, and the intensity of the intervention that was delivered (Borenstein et al., 2009). A random-effects model was utilized as the participants across studies varied in terms of their ethnicity, program, socioeconomic status, and location (see Table 1). However, as the decision to use one or the other model is subject to interpretation by the researcher, we include results for both models for the interested reader (see Tables 2 and 3).

Moderator analysis

The type of mathematics program, number of content strands, amount of time, type of learning environment, size of the group, and type of outcome measure were selected as moderator variables in

Table 2. Random-effects meta-analysis results for individual moderator variables.

Variable	No. of effects (Studies)	Mean effect size, Cohen's d (SE)	Effect size range	Homogeneity (Q)	Proportion of variance that is true (I^2)	ANOVA total between (Q_{TB})
Overall effect	31 (29)	0.62*** (0.06)	0.50–0.75	136.57***		
Moderator						
Program type						0.01
Comprehensive curriculum	10 (10)	0.63*** (0.10)	0.44–0.82	73.46***	88%	
Supplemental math-related activities	21 (19)	0.63*** (0.09)	0.45–0.81	62.86***	68%	
Content strand						0.93
Number and operations only	16 (14)	0.71*** (0.11)	0.48–0.93	48.92***	69%	
Multiple content strands	15 (15)	0.57*** (0.08)	0.41–0.72	83.18***	83%	
Average minutes per week when offered						0.39
23–60 min	14 (13)	0.60*** (0.06)	0.49–0.72	13.44	3%	
63–90 min	11 (10)	0.58*** (0.10)	0.39–0.77	57.15***	83%	
120–150 min	4 (4)	0.82** (0.29)	0.25–1.38	42.59***	93%	
Environment						0.35
Preschool	19 (19)	0.65*** (0.07)	0.51–0.78	58.85***	69%	
Kindergarten	12 (10)	0.60*** (0.12)	0.36–0.84	57.55***	81%	
Grouping						2.30
Include whole groups	10 (10)	0.51*** (0.09)	0.33–0.70	74.39***	88%	
Small groups	16 (14)	0.69*** (0.10)	0.49–0.90	44.52***	66%	
Individual only	5 (5)	0.74*** (0.15)	0.44–1.05	6.98	43%	
Type of test[a]						2.10
Standardized test	8 (8)	0.43*** (0.09)	0.26–0.60	9.26	24%	
Researcher-developed test	23 (22)	0.68*** (0.08)	0.53–0.83	124.49***	82%	

Note. ANOVA = analysis of variance.

[a] One study (Dyson et al., 2013) used both standardized and researcher-developed outcomes.

**p < .01. ***p < .001.

Table 3. Fixed-effects meta-analysis results for individual moderator variables.

Variable	No. of effects	Mean effect size, Cohen's d (SE)	Effect size range	Homogeneity (Q)	ANOVA total within (Q_{TW})	ANOVA total between (Q_{TB})
Overall effect	31 (29)	0.52*** (0.03)	0.47–0.57	136.57***		
Moderator						
Program type					136.31***	0.26
Comprehensive curriculum	10 (10)	0.53*** (0.03)	0.47–0.59	73.46***		
Supplemental math-related activities	21 (19)	0.50*** (0.05)	0.41–0.59	62.86***		
Content strand					132.09***	4.48*
Number and operations only	16 (16)	0.64*** (0.06)	0.52–0.76	48.92***		
Multiple content strands	15 (13)	0.49*** (0.03)	0.44–0.55	83.18***		
Average minutes per week when offered					113.18***	19.70***
23–60 min	14 (13)	0.60*** (0.06)	0.49–0.71	13.44		
63–90 min	11 (10)	0.57*** (0.03)	0.51–0.64	57.15***		
120–150 min	4 (4)	0.32*** (0.05)	0.22–0.42	42.59***		
Environment					116.40***	20.18***
Preschool	19 (19)	0.60*** (0.03)	0.54–0.66	58.85***		
Kindergarten	12 (10)	0.36*** (0.04)	0.28–0.45	57.55***		
Grouping					125.88***	10.69**
Include whole groups	10 (10)	0.48*** (0.03)	0.42–0.53	74.39***		
Small groups	16 (14)	0.65*** (0.06)	0.53–0.76	44.52***		
Individual only	5 (5)	0.73*** (0.11)	0.51–0.94	6.98		
Type of test[a]					133.75***	2.82
Standardized test	8 (8)	0.41*** (0.07)	0.27–0.55	9.26		
Researcher-developed test	23 (22)	0.54*** (0.03)	0.48–0.59	124.49***		

Note. ANOVA = analysis of variance.

[a] One study (Dyson et al., 2013) used both standardized and researcher-developed outcomes.

*p < .05. **p < .01. ***p < .001.

accordance with the purpose of this study. The type of mathematics program and the number of content strands varied across studies. For the moderator analysis, the type of mathematics program was defined as mathematics curriculum or supplemental mathematics-related activities. Studies that were coded as the math curriculum (n = 10) evaluated the effectiveness of math interventions that were implemented over an extended period of time with a specific scope and sequence. The 10 studies evaluated four curricula: Building Blocks Curriculum, Early Learning in Mathematics, Experimental Mathematics Curriculum, Pre-K Mathematics Curriculum. Studies that were coded as supplemental mathematics-related activities (n = 19) were less comprehensive interventions that were implemented in addition to the regular mathematics curriculum (see Tables 4 and 5).

The five mathematics content areas identified by the NCTM—number and operations, algebra, geometry, measurement, and data analysis and probability—served as a framework for classifying the content areas of the studies included in the current analysis, and each study was coded for the NCTM mathematical content area(s) that was the focus (see Table 4). For the moderator analysis, content was defined as focusing only on the mathematical content area of number and operations (n = 10) or focusing on multiple NCTM mathematical content areas, including number and operations (n = 15; see Table 4).

Table 4. Content strand by type of mathematics program.

Study	Number and operations	Algebra	Geometry	Measurement	Data analysis and probability
Curriculum studies (n = 10)					
Building Blocks					
Clements & Sarama (2007)	X		X		
Clements & Sarama (2008)	X	X	X	X	X
Clements et al. (2011)	X	X	X		
Sarama et al. (2008)[a]	X	X	X	X	
Sarama et al. (2012)	X		X		
Early Learning in Mathematics					
Chard et al. (2008)	X	X	X	X	
Clarke et al. (2011)	X		X	X	
Experimental Mathematics Curriculum					
Sophian (2004)	X		X	X	
Pre-K Mathematics Curriculum					
Klein (Starkey et al., 2008)	X	X	X	X	
Starkey et al. (2004)	X	X	X	X	
Supplemental math-related activities studies (n = 19)					
Arnold et al. (2002)	X				
Aunio et al. (2005)	X		X	X	
Baroody et al. (2009)	X				
Casey et al. (2008)			X		
Curtis et al. (2009)	X				
Dyson et al. (2013)	X				
Fuchs et al. (2001)	X			X	
Jordan (2012)	X				
Kidd et al. (2008)	X				
Monahan (2007)	X				
Pagani et al. (2006)	X	X	X	X	
Papic et al. (2011)	X	X			
Pasnak (2006)	X	X			
Ramani & Siegler (2008)	X				
Ramani & Siegler (2011)	X				
Siegler & Ramani (2009)	X				
Sood (2009)	X				
Tarim (2009)	X				
Young-Loveridge (2004)	X				

Note. Pre-K = prekindergarten.

[a] The mathematics curriculum used in the Sarama et al. (2008) study included two components: one from the Building Blocks (Clements & Sarama, 2007) project and the Pre-K Mathematics Curriculum (Klein et al., 2008).

Table 5. Summary of moderator variables.

Study	Subgroup within study	Cohen's d	Program type	Content strand	Grouping	Time	Time 2	Environment	Outcome type
Arnold et al. (2002)		0.40	Activities	One	Include whole group	85	63–90	Pre-K	Standardized (TEMA-2)
Aunio et al. (2005)		0.60	Activities	Multiple	Small group	60	23–60	Pre-K	Researcher developed (ENT)
Baroody et al. (2009)		0.53	Activities	One	Small group	75	63–90	Pre-K	Standardized (TEMA-3)
Casey et al. (2008)	Treatment 2	0.54	Activities	One	Small group	12	23–60	K	Researcher developed (Block Building/Mental Rotation)
Casey et al. (2008)	Treatment 2	0.70	Activities	One	Small group	12	23–60	K	Researcher developed (Block Building/Mental Rotation)
Chard et al. (2008)		0.32	Curriculum	Multiple	Include whole group	135	120–150	K	Standardized (SESAT-2)
Clarke et a. (2011)		0.22	Curriculum	Multiple	Include whole group	135	120–150	K	Researcher developed (CBM/TEMA-3)
Clements & Sarama (2008)		1.09	Curriculum	Multiple	Include whole group	63	63–90	Pre-K	Researcher developed (EMA)
Clements & Sarama (2007)		0.91	Curriculum	Multiple	Include whole group	63	63–90	Pre-K	Researcher developed (BBAEM)
Clements et al. (2011)		0.48	Curriculum	Multiple	Include whole group	63	63–90	Pre-K	Researcher developed (EMA)
Curtis et al. (2009)		0.82	Activities	One	Individual only	1	23–60	Pre-K	Researcher developed (Counting)
Dyson et al. (2013)	Outcome 1	0.13	Activities	One	Small group	90	63–90	K	Standardized (WJ App/Calc)
Dyson et al. (2013)	Outcome 2	0.42	Activities	One	Small group	90	63–90	K	Researcher developed (NSB)
Fuchs et al. (2001)		0.43	Activities	Multiple	Small group	40	23–60	K	Standardized (SESAT)
Jordan (2012)	Outcome 1	0.92	Activities	One	Small group	90	63–90	K	Standardized (WJ total)
Jordan (2012)	Outcome 2	1.11	Activities	One	Small group	90	63–90	K	Researcher developed (NSB)
Kidd et al. (2008)	Control 1	0.50	Activities	One	Small group	37.5	23–60	K	Standardized (OLSAT/WJ-III)
Kidd et al. (2008)	Control 2	0.61	Activities	One	Small group	37.5	23–60	K	Standardized (OLSAT/WJ-III)
Klein (Starkey et al., 2008)		0.52	Curriculum	Multiple	Small group	50	23–60	Pre-K	Researcher developed (CMA)
Monahan (2007)		0.25	Activities	One	Individual only	40	23–60	Pre-K	Researcher developed (ENCO)
Pagani et al. (2006)		0.26	Activities	Multiple	Include whole group	90	63–90	Pre-K	Researcher developed (NKT)
Papic et al. (2011)		0.30	Activities	Multiple	Small group	30	23–60	Pre-K	Researcher developed (EMPA)
Pasnak (2006)	Outcome 1	0.50	Activities	Multiple	Small group		Blank	Pre-K	Standardized (McCarthy/OLSAT)
Pasnak (2006)	Outcome 2	0.68	Activities	Multiple	Small group		Blank	Pre-K	Standardized (Oddity/Seriation)
Ramani & Siegler (2011)		1.39	Activities	One	Individual only	23	23–60	Pre-K	Researcher developed (Numeral Identification)
Ramani & Siegler (2008)		0.81	Activities	One	Individual only	34	23–60	Pre-K	Researcher developed (Numerical Magnitude)
Sarama et al. (2008)		0.46	Curriculum	Multiple	Include whole group	40	23–60	Pre-K	Researcher developed (REMA)
Sarama et al. (2012)		0.81	Curriculum	Multiple	Include whole group	78	63–90	Pre-K	Researcher developed (TEAM)

(*Continued*)

Table 5. (Continued).

Study	Subgroup within study	Cohen's d	Program type	Content strand	Grouping	Time	Time 2	Environment	Outcome type
Siegler & Ramani (2009)		0.94	Activities	One	Individual only	23	23–60	Pre-K	Researcher developed (Numerical Identification)
Sood (2009)		0.33	Activities	One	Include whole group	90	63–90	K	Researcher developed (EN-CBM)
Sophian (2004)		1.00	Curriculum	Multiple	Small group		Blank	Pre-K	Standardized (DSC)
Starkey et al. (2004)		0.84	Curriculum	Multiple	Small group	50	23–60	Pre-K	Researcher developed (CMA)
Tarim (2009)		1.08	Activities	One	Small group	120	120–150	K	Researcher developed (problem test)
Young-Loveridge (2004)		1.92	Activities	One	Small group	150	120–150	K	Researcher developed (individual task-based interviews)

Note. Pre-K = prekindergarten; K = kindergarten.

In addition, time, learning environment, size of the group, and type of outcome measure varied across studies (see Table 5). For the moderator analysis, time was defined and examined in two ways: as a continuous variable and as a categorical variable that captured the average number of minutes per week for the weeks when the early mathematics program was implemented; it ranged from 23 to 60 min per week (n = 13), to 63 to 90 min per week (n = 10), to 120 to 150 min per week (n = 4). For the practitioner, the categorical approach allows answers to questions such as "About how much time should be spent per week to be effective?" Time as a continuous variable representing the average number of minutes per week was created. Average time per week ranged from 1 to 150 min, with a mean of 64 min per week across all studies. Learning environment was defined as the pre-K environment (n = 19) or the kindergarten environment (n = 10). Size of the group was defined as including whole groups (n = 10), mainly small groups (n = 14), or individual only (n = 5). Type of outcome measure was defined as researcher-developed (n = 22) or standardized mathematics (n = 8) assessment because two studies used both types of assessments.

Although the moderator variables permitted a detailed examination of effect sizes according to key features of early childhood mathematics programs, imposing a moderator analysis on the small subset of studies eligible for this meta-analysis resulted in comparisons that were based on a small number of studies. Consequently, the results need to be interpreted cautiously. However, the small number of studies is a reflection of the current status of the literature to date and can hopefully direct the future study of early mathematics intervention.

Comprehensive Meta-Analysis Version 2.2.064 (Biostat, 2000) was used to calculate effect sizes (Cohen's d), examine moderator variables, and assess possible publication bias. A random-effects model was utilized in this meta-analysis as each study was assumed to have its own true effect size to reflect population differences in participant demographics and the intensity, focus, and nature of the intervention that was delivered (Borenstein et al., 2009). Under the random-effects model, more weight was given to studies that included more variance and larger sample sizes (Borenstein et al., 2009). We tested mean weighted effect sizes to discern whether they were greater than 0 at $p < .05$ in addition to testing the amount of variability in study effect sizes beyond sampling error by using Q statistics ($p < .05$). The existence of publication bias or reporting bias (Sterne et al., 2011) poses one of the greatest threats to the validity and credibility of meta-analytic results (Rothstein, Sutton, & Boreinstein, 2005). Publication bias exists when the included results of a meta-analytic search are systematically different from the population of completed research on a particular topic. The existence of bias can result in inaccurate, often overestimated effects of the intervention (Kepes, Banks, & Oh, 2014; Rothstein et al., 2005). The existence or absence of publication bias should be assessed using multiple methods using the triangulation process (Kepes et al., 2014; Rothstein et al., 2005). We assessed evidence of publication bias by using the funnel plot Begg and Mazumdar's rank correlation test, Egger's linear regression method, and Duval and Tweedie's trim-and-fill method (Borenstein et al., 2009).

Results

The overall effect size

A general effect size for the combined 29 studies of Cohen's $d = 0.62$ ($p < .001$) suggested a moderate to large effect for early childhood mathematics programs, following the convention that effect sizes of 0.20 or less are considered small, effect sizes of 0.50 are considered moderate, and effect sizes of 0.80 or larger are considered large. An effect size of 0.62 can be interpreted as suggesting that a child in the treatment group who performed at the mean of that group would fall at about the 74th percentile of the control group (Coe, 2012). Hence, the average treatment group child performed better than 74% of children in the control group. In addition, the results showed that there was significant heterogeneity around the average effect size, $Q(30) = 136.57$, $p < .001$. The moderator analyses reported here were designed to explain why average effect sizes varied across studies and programs. Table 2 provides an overview of the meta-analysis results, indicating the number of studies and effect sizes, mean effect size (Cohen's *d*), standard error and range, homogeneity statistic (Q), proportion of subgroup variance that reflected real differences in study effects (I^2), and meta-analysis of variance total between variance test (Q_{TB}). All reported mean effect sizes were statistically significant.

Supplemental activities versus comprehensive curricula

All 29 studies were included for this moderator analysis. A moderate effect size, 0.63 ($p < .001$), was noted for the 10 studies that reported on mathematics curriculum and for the 19 studies that reported on the mathematics-related activities. The test of heterogeneity for the effect size on mathematics curriculum studies was statistically significant, $Q(9) = 73.46$, $p < .001$, and the I^2 indicated that 88% of the subgroup variance reflected real differences in study effects. The test of heterogeneity for the effect size for the mathematics-related activities was statistically significant, $Q(20) = 62.86$, $p < .001$, and 68% of the subgroup variance reflected real differences in study effects. The two means suggested that the average child in the treatment group performed at about the 75th percentile of the control group.

Single versus multiple content strands

All 29 studies contributed to this moderator analysis. A moderate to large effect size, 0.71 ($p < .001$), was noted for the 14 studies that focused only on the mathematical content area number and operations. The test of heterogeneity for this effect size was statistically significant, $Q(13) = 48.92$, $p < .001$, and the I^2 indicated that 83% of the subgroup variance reflected real differences in study effects. A moderate effect size, 0.57 ($p < .001$), was noted for the 15 studies that focused on multiple mathematical content areas, including number and operations. The test of heterogeneity for this effect size was statistically significant, $Q(14) = 83.18$, $p < .001$, and 69% of the subgroup variance reflected real differences in study effects. Although the mean for programs that focused on a single content strand was higher than the mean for programs that focused on multiple strands, the confidence intervals for these mean effect sizes overlapped. Thus, this contrast between means was not statistically significant. A mean effect size of 0.71 suggests that the average child in the treatment group performed at about the 76th percentile of the control group. A mean effect size of 0.57 suggests that the average child in the treatment group performed at about the 72nd percentile of the control group.

Minutes per week

For this moderator analysis on minutes per week, 27 studies were included because two papers did not provide information about minutes per week. The authors of the two excluded studies were contacted, but they were not able to provide the necessary time information for this moderator analysis. A moderate effect size, 0.60 ($p < .001$), was noted for the 13 studies that presented math programs 23 to 60 min per week on average during the weeks when math was offered. This finding suggests that the average child in the treatment group performed better than 73% of children in the control group. The test of heterogeneity for this effect size was not significant, $Q(11) = 13.44$, $p > .05$. Similarly, a moderate effect size, 0.58 ($p < .001$), was found for the 10 studies that presented math programs 63 to 90 min per week on average. This finding suggests that the average child in the treatment group performed at about the 72nd percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(10) = 57.15$, $p < .001$, and the I^2 indicated that 83% of the subgroup variance reflected real differences in study effects. In contrast, a large effect size of 0.82 ($p < .01$) was reported for the four studies that presented math programs 120 to 150 min per week on average. This finding suggests that the average child in the treatment group performed at about the 78th percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(3) = 42.59$, $p < .001$, and the I^2 indicated that 93% of the subgroup variance reflected real differences in study effects. Although all of the mean effect sizes were significantly larger than zero, none of these contrasts between means were statistically significant because the confidence intervals for these means overlapped.

In addition to conducting a categorical analysis for the time variable, we also conducted a meta-regression analysis treating time as a continuous variable using Comprehensive Meta-Analysis (Borenstein et al., 2009). The Q value for the model was $Q(1) = 4.21$, $p < .05$, which indicated that effect size was related to time. However, the regression coefficient for time as a continuous measure of average minutes per week was $B = -0.003$ ($SE = 0.002$, $p < .05$), which implied that an intervention that was offered 100 min per week had on average a 0.3 higher effect size than an intervention of 200 min per week.

Programs for pre-K environments versus kindergarten

All 29 studies contributed to this moderator analysis. A moderate effect size, 0.65 ($p < .001$), was noted for the 19 studies that focused on mathematics programs for pre-K environments. This means the average treatment group child performed at the 75th percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(18) = 58.85$, $p < .001$, and the I^2 indicated that 69% of the subgroup variance reflected real differences in study effects. A moderate effect size, 0.60 ($p < .001$), was found for the 10 studies that focused on mathematics programs for kindergarten. This means the average treatment group child performed at the 73rd percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(11) = 57.55$, $p < .001$, and the I^2 indicated that 81% of the subgroup variance reflected real differences in study effects. This contrast between the mean effect sizes was not statistically significant because the confidence intervals for these means overlapped.

Individual versus small-group versus whole-class format

For the moderator analysis on grouping strategies that were used, all 29 studies were included. A moderate effect size 0.51 ($p < .001$) was noted for the 10 studies that used whole groups. This means the average child in the treatment group performed at the 69th percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(9) = 74.39$, $p < .001$, and the I^2 indicated that 88% of the subgroup variance reflected real differences in study effects. Similarly, a moderate to large effect size, 0.69 ($p < .05$), was noted for the 14 studies that used mainly small groups. This means the average child in the treatment group performed at the 76th percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(15) = 44.52$, $p < .001$, and the I^2 indicated that 66% of the subgroup variance reflected real differences in study effects. A moderate to large effect size, 0.74 ($p < .001$), was noted for the five studies that used individuals only. This means the average child in the treatment group performed at the 78th percentile of the control group. The test of heterogeneity for this effect size was not significant, $Q(4) = 6.98$, $p > .05$, which was supported by the small I^2. Although all of the mean effect sizes were significantly larger than zero, none of these contrasts between means were statistically significant because the confidence intervals for these means overlapped.

Researcher-made versus standardized outcome measures

All 29 studies contributed to this moderator analysis. A small to moderate effect size, 0.43 ($p < .001$), was noted for the eight studies that used standardized outcome measures. This means that the average child in the treatment group performed at the 67th percentile of the control group. The test of heterogeneity for this effect size was not statistically significant, $Q(7) = 9.26$, $p > .05$, as supported by the small I^2. A moderate to large effect size, 0.68 ($p < .001$), was found for the 22 studies that used researcher-made outcome measures. This means that the average child in the treatment group performed at the 75th percentile of the control group. The test of heterogeneity for this effect size was statistically significant, $Q(22) = 124.49$, $p < .001$, and the I^2 indicated that 82% of the subgroup variance reflected real differences in study effects. As the confidence intervals for these mean effect sizes overlapped, this contrast was not statistically significant. Also, an additional moderator analysis was conducted to see whether there was a significant difference in mean effect size between studies that used experimental designs ($d = 0.64$, $n = 22$) and those that used quasi-experimental designs ($d = 0.58$, $n = 7$); the difference was not statistically significant, $Q_{\text{TotalBetween}}(1) = 0.13$, $p > .05$.

Publication bias

With any meta-analysis, one has to be concerned that only the largest effects were published and that there may be studies that found no difference that were not published. We assessed the existence of such a publication bias using multiple methods. We first conducted a review of the funnel plot (Rothstein e al., 2005), or a scatterplot of the magnitude of effect sizes against the standard error, to determine initial evidence of publication bias by looking for an asymmetrical distribution of studies around the mean effect size (Kepes et al., 2014). A symmetrical distribution of studies indicates the absence of bias, whereas an asymmetrical distribution indicates potential bias. Our funnel plot revealed relative symmetrical distribution of the large studies around the mean effect size, although there were two outliers that appeared on the right side of the mean effect size.

The presence of bias was then statistically tested using several tests, including Begg and Mazumdar's (1994) rank correlation test and Egger's test of the intercept (Egger, Smith, Schneider, & Minder, 1997). Nonsignificant rank-order correlation indicates the absence of bias, whereas significant rank-order correlation indicates the existence of bias. Our rank correlation test demonstrated an absence of bias with a nonsignificant rank-order correlation (Kendall's $\tau b = 0.18$, $p > .05$, one-and two-tailed). Egger's test of the intercept partially supported the absence of bias with a nonsignificant intercept at the .05 level, two-tailed test ($B_0 = 1.28$), $t(29) = 1.96$, $p < .05$ (one-tailed; $p > .05$, two-tailed).

Furthermore, we assessed the degree of symmetry in the funnel plot and adjusted the meta-analytic results for publication bias by using the recommended Duval and Tweedie (2000a, 2000b) trim-and-fill method. Rothstein et al. (2005) suggested that publication bias is likely to be absent or negligible when the meta-analytic trim-and-fill imputed effect size estimates do not differ much (or less than 20%) from the point estimate. The difference between the trim-and-fill imputed point estimate (0.48) and the random-effects model point estimate (0.62) was 15%. The difference between the trim-and-fill imputed point estimate (0.46) and the fixed-effects model point estimate (0.52) was 6%. The size of these differences indicated absent or negligible publication bias.

In summary, by triangulating results from the funnel plot, the Begg and Mazumdar (1994) rank correlation test, Egger's test of the intercept, and Duval and Tweedie (2000b, 2000b) trim-and-fill method, we were able to determine that there was an absence of publication bias, providing validity and credibility for our meta-analytic results and the means we report.

Discussion

The primary purpose of the present meta-analysis was to extend beyond prior evaluations of early childhood programs (e.g., Frye et al., 2013) by quantifying program effectiveness in terms of effect sizes and examining six aspects of programs that have both practical and theoretical value. Our results showed that the average effect size across all programs was moderate to large ($d = 0.62$) and that there was a tendency for programs to produce larger effects when they did the following: (a) targeted a single content strand rather than multiple content strands, (b) presented 120 to 150 min per week compared to 23 to 60 min per week or 63 to 90 min per week, (c) designed programs for the pre-K environment compared to the kindergarten environment, (d) presented content individually to children as opposed to small groups or to the entire class, and (e) used researcher-made compared to standardized mathematics assessments. In what follows, we consider the practical and theoretical implications of each of these findings in turn.

The overall effect size

The fact that the overall average effect size for all programs was moderate (d = 0.62) implies that interventions designed to improve the mathematics skills of young children can clearly have an effect and are not limited to a single setting or particular program. Although it would be reasonable to surmise that this strong effect derives from the fact that so little time is spent on math in early childhood programs (and presenting any content is a big improvement on presenting very little or no content), the overall average effect size reported here is comparable in size to what has been found for early phonological awareness training programs (National Institute of Child Health and Human Development, 2000) and larger than what has been found for alphabet training programs for young children (Piasta & Wagner, 2010). Thus, the findings suggest that the programs are generally well designed and implemented effectively to have such an effect. Moreover, it would clearly be worth the effort to implement mathematics interventions on a large scale nationally, given the importance of mathematics skills and the need to begin early. There was, however, significant heterogeneity around the mean value, suggesting that some programs were more effective than others. In particular, about 40% of the effects ranged from d = 0.22 to d = 0.48, about 20% ranged from d = 0.52 to d = 0.62, and about 40% ranged from d = 0.81 to d = 1.92. We sought to explain such variability by considering six aspects of these programs, which we discuss next.

Supplemental activities versus comprehensive curricula

The fact that the average effect size for programs that were designed to be supplemental add-ons to existing curricula was comparable to that for more comprehensive, stand-alone curricula that would replace existing programs seems to suggest that the most practical conclusion for early childhood educators is to opt for the supplemental add-ons given the time, expense, and professional development needed to learn a new curriculum. However, of the three programs that had the strongest effects (>1.0 effect sizes), one involved using books to engender numeracy (Young-Loverage, 2004), one was a comprehensive curriculum (i.e., Building Blocks of Clements and Sarama), and one was a board game (Siegler & Ramani, 2009). So it clearly matters which activities or comprehensive curricula are at issue. Moreover, it is important to stress that most outcome measures only assessed mathematics knowledge and skills rather than other factors, such as interest in mathematics and whether children mathematize their environments (Clements & Sarama, 2009). Longitudinal studies of early mathematics skills have shown that motivational factors add significant variance toward predicting mathematics achievement outcomes over and above prior achievement (e.g., Byrnes & Wasik, 2009). In addition, few studies examined the long-term effects of these programs after they were terminated. It is possible that the more comprehensive programs would have enduring effects. Such proposals should be examined in future studies.

However, it is also possible that supplemental, game-like activities that target key, foundational abilities (e.g., numerosities and the ordinal representation of magnitude lines) have effects that generalize (Siegler & Ramani, 2008). Activities that simply present content repetitively via cards and do not target such foundational activities seemed to have a weaker effect (e.g., Chard et al., 2008, d = 0.32). This proposal should be tested experimentally in follow-up experiments.

Single versus multiple content strands

Psychological research on memory (Byrnes, 2008) suggests that, all things being equal, programs that target a single content strand of mathematics (e.g., number and operations) would produce stronger effects than programs that target multiple content strands (e.g., number and operations, geometry and patterns) because a singular focus allows more time and practice for that strand. The results revealed a tendency toward this hypothesis (i.e., Number and operations only, d= 0.71, vs. Multiple content strands, d= 0.57), but the contrast was not significant. However, as noted earlier, this expectation would only be borne out if outcome measures matched the focus of the program. If outcome measures focused on five content strands, it would be expected that the multiple-strand programs would have stronger effects than the single-strand programs. In addition, each of the two program types would have to be implemented for the same amount of time each week. Thus, a true test of the proposal requires that researchers conduct additional studies with common outcome measures and common dosages. In addition, however, part of the reason for the nonsignificance of the contrast could be the low sample sizes and large standard deviations in some of the studies, which produced large confidence intervals around each mean (that overlapped).

Minutes per week

Again, all things being equal, it seems reasonable to expect that programs that spend more time per week presenting mathematics content would produce stronger effects than programs that spend less time. As noted previously, however, there was a tendency for programs that infused mathematics activities for 120 to 150 min per week to have stronger effects than those that spent less time, but the contrasts between the three time levels (i.e., 23–60 min, 63–90 min, 120–150 min) were not significant. There are four ways to explain this finding. The first is that there may be diminishing returns from content exposure in learning situations. That is, beyond a certain point additional practice or exposure only has minimal effects (Anderson, 2009). However, this explanation requires that more time = more practice. The authors of the studies reviewed in this meta-analysis did not provide enough details on their programs for us to determine whether this equation held. It is possible that some activities took more time but did not provide additional practice of the content.

The second explanation is that this finding is similar to a growing number of findings in the area of language development, in which teachers' more frequent use of vocabulary words does not necessarily lead to growth in children's vocabulary (Rowe, 2012; Wasik & Hindman, in press). In other words, it is not how often a word is used but when and how—the quality (or manner) matters more than the quantity. Again, of the three programs that produced the strongest effects, one was in the top category for minutes per week (Young-Loveridge, 2004), one was in the middle category (Clements & Sarama, 2007), and one was in the lowest category for time (Ramani & Siegler, 2011).

The third explanation derives from the opportunity-propensity model of achievement (e.g., Byrnes & Wasik, 2009; Wang et al., 2013). In this model, learning will only occur if children have the propensity to take advantage of the learning opportunities presented to them (in this case by early mathematics programs). By *propensity*, it is meant that students have the requisite prior knowledge in mathematics and motivation to learn the content. A time variable would predict learning when propensity factors are taken into account. This explanation would be tested by predicting end-of-program mathematics scores using initial mathematics knowledge, motivation for mathematics, and instructional condition (intervention, control) as predictors.

The fourth explanation was mentioned earlier: There was a tendency for programs that were implemented for more time to have stronger effects (120–150 min: d=0.82, 63–90 min: d=0.58, and 23–60 min: d=0.60), but the contrasts were not significant because the confidence intervals overlapped. Confidence intervals are a function of the sample sizes of studies and standard deviations around means. It is possible that the contrasts would become significant if additional studies were conducted with larger sample sizes and more precise measures.

Programs for pre-K environments versus kindergarten

The effect sizes for pre-K environments and kindergarten were nearly identical (d = 0.65 and 0.60, respectively). This finding is somewhat surprising given the differing cognitive, self-regulatory, and motivational profiles of these two age groups as well as different kinds of structure in the two environments. However, if researchers took these differences into account and designed developmentally appropriate programs, comparable effect sizes would be expected. The present results suggest that researchers did just that.

Individual versus small-group versus whole-class format

Although average effect sizes for the three formats were aligned with expectations (individual > small group > whole class), the contrasts were not significant. The most likely explanation for this result has to do with the number of studies, samples sizes, and variance in the outcome measures. We suspect that when additional studies are conducted, the expected differences will emerge because the confidence intervals around the three average effect sizes will become narrower.

Researcher-made versus standardized outcome measures

As might be expected from the fact that researchers designed outcome measures to match the content of their programs, there was a tendency for effect sizes to be larger when researcher-made assessments were used (d = 0.68) than when standardized measures such as the Test of Early Mathematics Ability were used (d = 0.43). However, once again this contrast was not significant because of overlapping confidence intervals. Conducting additional studies with larger sample sizes should remedy this situation, but it would be useful for researchers to utilize common measures (rather than their own measures) in order to truly evaluate the relative effectiveness of their respective programs. If all were to use their own measures that closely match the skills taught, a comparison of effect sizes would not be particularly meaningful (e.g., a d = 0.50 using Measure A would not necessarily be comparable to a d = 0.50 for Measure B). The goal would be to use measures that align with high-stakes tests or outcomes valued by key constituencies, so that apples could be compared to apples, so to speak.

Limitations and future directions

Similar to other meta-analyses, there are a few common limitations to the current meta-analysis. These include the potential problem of generalizing results from the small set of available studies, overreliance on studies that were published, and unavailability of information on key program features in published and unpublished studies.

Specific to this meta-analysis is the possible limitation of generalizability due to the small number of quasi-experimental and experimental studies found in this area. We found only 29 studies in the past 15 years that focused on the outcome of early mathematics programs for pre-K or kindergarten students, utilized randomized controlled trials or quasi-experimental designs with treatment and control groups, and provided enough information on different program features. We were surprised to have found so few after our exhaustive search process, especially with the number of early childhood mathematics programs that have been created over the past 15 years. Given how critical the first 5 years of life are in developing children’s early mathematics skills (Byrnes & Wasik, 2009;

Duncan et al., 1994; National Institute of Child Health and Human Development Early Child Care Research Network, 2006; Wang et al., 2013), we urge researchers in the field to continue conducting studies that will provide sound evidence on developmentally appropriate and highly effective early mathematics programs.

Another potential limitation on generalizability is posed by the overreliance on published studies. In our search process, we included studies from a range of publication channels by searching for studies appearing in academic journals, books, or conference proceedings; those released by an organization; and published dissertations. Although we tested for the presence of publication bias using a wide variety of techniques and found minimal to no evidence of publication bias, nevertheless the possibility of publication bias is still present. We urge meta-analysis researchers who have more resources and time to devote a portion of their resources to the development and implementation of an ambitious search process for dissertations and other unpublished work to address this limitation. In addition, there is emerging evidence that meta-analysts might want to report random-effects, fixed-effects, and unrestricted weighted least squares estimates when key policy parameters are involved (Stanley & Doucouliagos, 2015).

Moreover, in a couple instances, studies included in the meta-analysis did not report enough information for the moderator variables of interest, and when contacted a few authors could not provide such information. The inability to use all studies in each moderator analysis could have decreased the statistical power of the meta-analysis. This suggests that all published and unpublished studies should be encouraged to include detailed information related to the type of mathematics program, number and type of content strands, amount of time devoted to mathematics activities, type of learning environment, size of the group, and type of outcome measure, in addition to the demographics of the sample, the procedure of the study, and the outcomes of the study. Furthermore, as a field, more experimental studies on interventions created based on well-researched psychological basis need to be conducted in early childhood education context so that in the future, meta-analysts like us can test theory-based hypotheses, similar to what De Boer, Donker, and Van Der Werf (2014) were able to do in their meta-analysis of 58 studies and 95 interventions published between 2000 and 2011 on primary and secondary students' academic performance.

Conclusions

Once again, it is clear that early mathematics programs can be designed that are both developmentally appropriate and highly effective. The goal of this meta-analysis was to reveal these programs and some of the factors that might explain why some produced stronger effects than others. It is hoped that researchers and policymakers can use our suggestions for future research to firm up recommendations and understand the kinds of studies that are still needed.

References

References marked with an asterisk were included in the meta-analysis.

Anderson, J. R. (2009). *Cognitive psychology and its implications* (7th ed.). New York, NY: Worth.

*Arnold, D. H., Fisher, P. H., Doctoroff, G. L., & Dobbs, J. (2002). Accelerating math development in Head Start classrooms. *Journal of Educational Psychology*, *94*(4), 762–770. doi:10.1037/0022-0663.94.4.762

*Aunio, P., Hautamäki, J., & Van Luit, J. E. H. (2005). Mathematical thinking intervention programmes for preschool children with normal and low number sense. *European Journal of Special Needs Education*, *20*(2), 131–146. doi:10.1080/08856250500055578

*Baroody, A. J., Eiland, M., & Thompson, B. (2009). Fostering at-risk preschoolers' number sense. *Early Education & Development*, *20*(1), 80–128. doi:10.1080/10409280802206619

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

Biostat. (2000). *Comprehensive meta-analysis* (Version 2.2) [Software]. Englewood, NJ: Author.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (n.d.). *Comprehensive meta-analysis* (Version 2.0). Englewood, NJ: Biostat.

Byrnes, J. P. (2008). *Cognitive development and learning in instructional contexts* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Byrnes, J. P., & Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grades: An opportunity-propensity analysis. *Contemporary Educational Psychology*, *34*, 167–183. doi:10.1016/j.cedpsych.2009.01.002

*Casey, B. M., Andrews, N., Schindler, H., Kersh, J. E., Samper, A., & Copley, J. (2008). The development of spatial skills through interventions involving block building activities. *Cognition and Instruction*, *26*(3), 269–309. doi:10.1080/07370000802177177

*Chard, D. J., Baker, S. K., Clarke, B., Jungjohann, K., Davis, K., & Smolkowski, K. (2008). Preventing early mathematics difficulties: The feasibility of a rigorous kindergarten mathematics curriculum. *Learning Disability Quarterly*, *31*(1), 11–20.

*Clarke, B., Smolkowski, K., Baker, S., Fien, H., Doabler, C., & Chard, D. (2011). The impact of a comprehensive Tier I core kindergarten program on the achievement of students at risk in mathematics. *The Elementary School Journal*, *111*(4), 561–584. doi:10.1086/659033

*Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163.

*Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443–494. doi:10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach*. New York, NY: Taylor & Francis.

*Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166.

Coe, R. J. (2012). Effect size. In J. Arthur, M. Waring, R. Coe, & L. V. Hedges (Eds.), *Research methods and methodologies in education* (pp. 368–377). Thousand Oaks, CA: Sage.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *Handbook of research synthesis*. New York, NY: Russell Sage Foundation.

*Curtis, R., Okamoto, Y., & Weckbacher, L. M. (2009). Preschoolers' use of count information to judge relative quantity. *Early Childhood Research Quarterly*, *24*(3), 325–336. doi:10.1016/j.ecresq.2009.04.003

De Boer, H., Donker, A. S., & Van Der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research*, *84*(4), 509–545. doi:10.3102/0034654314540006

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development*, *65*, 296–318. doi:10.2307/1131385

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

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

*Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*(2), 166–181. doi:10.1177/0022219411410233

Early, D. M., Barbarin, O., Bryant, D., Burchinal, M., Chang, F., Clifford, R., Crawford, G., & Weaver, W. (2005). *Prekindergarten in eleven states: NCEDL's Multi-State Study of Pre-Kindergarten and Study of State-Wide Early Education Programs (SWEEP)*. Chapel Hill, NC: Frank Porter Graham Child Development Institute.

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

Fisher, K. R., Hirsh-Pasek, K., Newcombe, N., & Golinkoff, R. M. (2013). Taking shape: Supporting preschoolers' acquisition of geometric knowledge through guided play. *Child Development*, *84*, 1872–1878. doi:10.1111/cdev.2013.84.issue-6

Frye, D., Baroody, A. J., Burchinal, M., Carver, S. M., Jordan, N. C., & McDowell, J. (2013). *Teaching math to young children: A practice guide* (NCEE 2014-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance.

*Fuchs, L. S., Fuchs, D., & Karns, K. (2001). Enhancing kindergartners' mathematical development: Effects of peer-assisted learning strategies. *The Elementary School Journal*, *101*(5), 495–510. doi:10.1086/esj.2001.101.issue-5

Hausken, E. G., & Rathbun, A. (April, 2004). *Mathematics instruction in kindergarten: Classroom practices and outcomes*. San Diego, CA: American Educational Research Association.

Hedges, L. V., & Stock, W. (1983). The effects of class size: An examination of rival hypotheses. *American Educational Research Journal*, *20*, 63–85. doi:10.3102/00028312020001063

*Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergarteners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647–660. doi:10.1037/a0029018

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

*Kidd, J. K., Pasnak, R., Gadzichowski, M., Ferral-Like, M., & Gallington, D. (2008). Enhancing early numeracy by promoting abstract thought involved in the oddity principle, seriation, and conservation. *Journal of Advanced Academics*, *19*(2), 164–200.

*Klein, A., Starkey, P., Clements, D. H., Sarama, J., & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. *Journal of Research on Educational Effectiveness*, *1*(3), 155–178. doi:10.1080/19345740802114533

Legnard, D. S., & Austin, S. L. (2012). The menu for every young mathematician's appetite. *Teaching Children Mathematics*, *19*(4), 228–236. doi:10.5951/teacchilmath.19.4.0228

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

Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., & Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interventions: To what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children? *Early Childhood Research Quarterly*, *25*(4), 409–431. doi:10.1016/j.ecresq.2010.03.002

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). *Quantitative development in infancy and early childhood*. New York, NY: Oxford University Press.

*Monahan, S. (2007). *Emergent Numeracy and Cultural Orientations (ENCO) project: Examining approaches to meaningful and contextual mathematics instruction*. Dissertations available from ProQuest. Paper AAI3271792. http://repository.upenn.edu/dissertations/AAI3271792

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386. doi:10.1177/1094428106291059

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: Boston College.

National Association for the Education of Young Children. (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8*. Washington, DC: Author.

National Center for Education Statistics. (2011). *The nation's report card: Mathematics 2011*. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012458

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through Grade 8 mathematics: A quest for coherence*. Reston, VA: Author.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

National Institute of Child Health and Human Development Early Child Care Research Network. (2006). Child-care effect sizes for the NICHD Study of Early Child Care and Youth Development. *American Psychologist*, *61*, 99–116. doi:10.1037/0003-066X.61.2.99

National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity* (C. T. Cross, T. A. Woods, & H. Schweingruber, Eds.). Washington, DC: National Academies Press.

*Pagani, L. S., Jalbert, J., & Girard, A. (2006). Does preschool enrichment of precursors to arithmetic influence intuitive knowledge of number in low income children? *Early Childhood Education Journal*, *34*(2), 133–146. doi:10.1007/s10643-005-0034-2

*Papic, M. M., Mulligan, J. T., & Mitchelmore, M. C. (2011). Assessing the development of preschoolers' mathematical patterning. *Journal for Research in Mathematics Education*, *42*(3), 237–269.

*Pasnak, R. (2006). Applying principles of development to help at risk preschoolers develop numeracy. *Journal of Psychology*, *140*(2), 155–173. doi:10.3200/JRLP.140.2.155-173

Peat, J. (2001). *Health science research: A handbook of quantitative methods*. Sydney, Australiabk_AQCmts13b: Allen & Unwin.

Piasta, S. B., & Wagner, R. K. (2010). Developing early literacy skills: A meta-analysis of alphabet learning and instruction. *Reading Research Quarterly*, *45*, 8–38. doi:10.1598/RRQ.45.1.2

*Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, *79*(2), 375–394. doi:10.1111/j.1467-8624.2007.01131.x

*Ramani, G. B., & Siegler, R. S. (2011). Reducing the gap in numerical knowledge between low- and middle-income preschoolers. *Journal of Applied Developmental Psychology*, *32*(3), 146–159. doi:10.1016/j.appdev.2011.02.005

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

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*, 1762–1774. doi:10.1111/j.1467-8624.2012.01805.x

*Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*(2), 89–119. doi:10.1080/19345740801941332

*Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*, 489–502. doi:10.1016/j.ecresq.2011.12.002

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science*, *11*, 655–661. doi:10.1111/desc.2008.11.issue-5

*Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology*, *101*, 545–560. doi:10.1037/a0014239

*Sood, S. (2009). *Teaching number sense: Examining the effects of number sense instruction on mathematics competence of kindergarten students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3373089)

*Sophian, C. (2004). Mathematics for the future: Developing a Head Start curriculum to support mathematics learning. *Early Childhood Research Quarterly*, *19*(1), 59–81. doi:10.1016/j.ecresq.2004.01.015

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine*, *34*, 2116–2127. doi:10.1002/sim.6481

*Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly*, *19*(1), 99–120. doi:10.1016/j.ecresq.2004.01.002

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal*, *342*, d4002. doi:10.1136/bmj.d4002

*Tarim, K. (2009). The effects of cooperative learning on preschoolers' mathematics problem-solving ability. *Educational Studies in Mathematics*, *72*(3), 325–340. doi:10.1007/s10649-009-9197-x

Van De Walle, J. A., Lovin, L. A. H., Karp, K. S., & Bay-Williams, J. M. (2013). *Teaching student-centered mathematics: Developmentally appropriate instruction for Grades pre-K to 2* (2nd ed.). Upper Saddle River, NJ: Pearson.

Wang, A. H. (2010). Optimizing early mathematics experiences for children from low-income families: A study on opportunity to learn mathematics. *Early Childhood Education Journal*, *37*, 295–302. doi:10.1007/s10643-009-0353-9

Wang, A. H., Shen, F., & Byrnes, J. (2013). Does the opportunity–propensity framework predict the early mathematics skills of low-income pre-kindergarten children? *Contemporary Educational Psychology*, *38*, 259–270. doi:10.1016/j.cedpsych.2013.04.004

Wasik, B. A., & Hindman, A. H. (in press). Closing the thirty million word gap: Why talking more is not enough. *Phi Delta Kappan*.

Wilson, D. B. (n.d.). *Practical meta-analysis effect size calculator*. Retrieved from http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD28.php

*Young-Loveridge, J. (2004). Effects on early numeracy of a program using number books and games. *Early Childhood Research Quarterly*, *19*(1), 82–98. doi:10.1016/j.ecresq.2004.01.001