AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–19
DOI: 10.1177/23328584211004183
Article reuse guidelines: sagepub.com/journals-permissions

Measures Matter: A Meta-Analysis of the Effects of Educational Apps on Preschool to Grade 3 Children's Literacy and Math Skills

James Kim
Harvard University Graduate School of Education

Joshua Gilbert
Harvard University Graduate School of Education
New England Conservatory

Qun Yu
Boston College

Charles Gale
Harvard University Graduate School of Education

Thousands of educational apps are available to students, teachers, and parents, yet research on their effectiveness is limited. This meta-analysis synthesized findings from 36 intervention studies and 285 effect sizes evaluating the effectiveness of educational apps for preschool to Grade 3 children and the moderating role of methodological, participant, and intervention characteristics. Using random effects meta-regression with robust variance estimation, we summarized the overall impact of educational apps and examined potential moderator effects. First, results from rigorous experimental and quasi-experimental studies yielded a mean weighted effect size of +0.31 standard deviations on overall achievement and comparable effects in both math and literacy. Second, the positive overall effect masks substantial variability in app effectiveness, as meta-regression analyses revealed three significant moderators of treatment effects. Treatment effects were larger for studies involving preschool rather than K–3 students, for studies using researcher-developed rather than standardized outcomes, and for studies measuring constrained rather than unconstrained skills.

Keywords: *educational apps, meta-analysis, literacy, math, academic achievement, constrained and unconstrained skills, preschool, early elementary*

Digital educational applications ("apps") are an increasingly appealing tool for promoting young children's school readiness and basic literacy and math skills. In particular, apps that run on touchscreen tablets and smartphones are now a ubiquitous feature of children's homes and schools. For example, a recent study on app usage in schools noted that there are over 2,500 education apps available to school leaders (S. Baker & Gowda, 2018), and the market for educational software is estimated in the billions of dollars in the United States (Richards & Stebbins, 2014). Similarly, parents are now confronted with an ever-increasing number of apps to improve children's academic achievement; the number of educational and reference apps in Apple's App Store has increased from 80,000 in 2015 to 200,000 in 2018 (Hirsh-Pasek et al., 2015; Pendlebury, 2018). More recently, the spread of the COVID-19 pandemic has ignited efforts by research and policy organizations to offer free and easy-to-use educational apps as a scalable strategy for helping young children acquire and maintain basic literacy and mathematics skills (U.S. Department of Education, 2020).

Despite the proliferation of educational apps designed for young children from preschool to Grade 3, effectiveness research on the causal impact of educational apps is in its infancy. Reviewing research on school-based educational apps, Haßler et al. (2016) concluded that "the fragmented nature of the current knowledge base, and the scarcity of rigorous studies, makes it difficult to draw firm conclusions" (p. 139). More specifically, because children use apps in

diverse ways from watching YouTube, to browsing the Internet, to playing video games (Radesky et al., 2020; Xie et al., 2018), and for a variety of other purposes, rigorous experimental designs are needed to isolate the causal effects of educational apps. Research over the past decade has focused on the potential and pitfalls of the medium—that is, touchscreen technologies—rather than the content and quality of activities on interactive apps (Madigan et al., 2019; Wexler, 2019; World Health Organization, 2019).

This meta-analytic review focuses on a specific type of intervention—namely, educational apps designed to improve the literacy and mathematics skills of preschool to third-grade children—in order to quantify mean effects and to identify factors that may enhance or diminish their effectiveness (Guernsey et al., 2012; Haßler et al., 2016; Papadakis et al., 2018). Given the proliferation of apps targeting children ages 3 to 9 (American Academy of Pediatrics, 2016) and the importance of building foundational literacy and math skills necessary for future academic success (National Research Council, 2015; Yoshikawa et al., 2016), our review focused on studies of educational apps for preschool to Grade 3.

Defining "Educational Apps"

It is critical to define the term *educational app* because it has been used inconsistently in the broader research literature. In this review, educational apps are defined as interventions designed to improve prekindergarten through third-grade children's literacy and mathematics skills (Cherner et al., 2014; Notari et al., 2016) through content delivered on smart phones, tablets, or personal computers (Hirsh-Pasek et al., 2015). Skill-building apps comprise the largest group of apps in the marketplace (Notari et al., 2016) and can be clearly distinguished from apps with other goals, including collaboration apps, learning and teaching support apps for instructors, communication apps, and reference apps. Therefore, our review of educational apps excludes eBooks; content-based apps that provide information like maps or dictionaries; function-based apps that provide tools for presentations, communication, and collaboration (Cherner et al., 2014); and apps that target domains outside of literacy and math, such as social-emotional skills, social studies, or science.

Within the academic domains of literacy and math, educational apps can also target improvement in constrained or unconstrained skills from preschool to third grade (Lipsey et al., 2018; McCormick et al., 2020; Paris, 2005; Snow & Matthews, 2016). Constrained skills are often more sensitive to direct teaching interventions, have a ceiling, and are mastered by most children. For example, one-to-one tutoring, small-group instruction, and whole-classroom interventions typically have their largest impact on constrained skills such as letter knowledge, print awareness, and phonemic awareness in literacy and counting, sorting shapes, and simple sums in math (Pearson et al., 2020; Wong et al., 2008). In contrast, unconstrained skills include broader domains of knowledge and include outcomes like math problem solving and vocabulary.

What Is Known About the Effectiveness of Educational Apps?

Although children are spending more time on educational apps in both school and home contexts (Rideout & Robb, 2020), there is surprisingly little causal evidence about their effectiveness or the features that enhance or diminish their effectiveness. To date, there is mixed evidence that educational apps improve student outcomes. Although there is some evidence that educational apps can improve early-grade math skills (Schaeffer et al., 2018), a narrative review of apps for preschool-aged children concluded that "more large-scaled randomized trials of apps are needed" (Griffith et al., 2020, p. 11). One way to synthesize the existing research with timely and rigorous evidence is to use meta-analytic methods to combine results from small- to medium-sized experiments and quasi experiments and to explore potential sources of treatment effect heterogeneity.

During the past 5 years, scholars in diverse fields such as developmental pediatrics, cognitive psychology, educational technology, and early education have published reviews of educational apps. As shown in Table 1, none of these previous review studies have attempted to conduct a meta-analysis that combines effect sizes from intervention studies or to explore how intervention, participant, or methodological factors explain variation in effects. A consistent conclusion in all the reviews is the need for more randomized experimental designs that provide stronger causal evidence regarding the effectiveness of educational apps and examination of the factors that moderate the effectiveness of educational apps on young children's learning (Griffith et al., 2020; Hainey et al., 2016; McTigue et al., 2020).

Research Questions and Hypotheses

Both theoretical and empirical research drawn from the science of learning suggest interactive educational applications can support active, engaging, targeted, and varied practice (Bjork, 1994; Griffith et al., 2020; Hirsh-Pasek et al., 2015; Pashler et al., 2007). There are several potential mechanisms through which educational apps may improve student learning, including the medium, the context, and the affordances of gamified learning. First, touchscreen technologies do not require young children to have the fine-motor skills needed to use computer keyboards and the mouse (Flewitt et al., 2015; Kucirkova, 2014), making them an engaging medium and easy-to-use technology for young children. Second, educational apps are typically employed in one-to-one or small-group contexts

TABLE 1
Findings From Recent Reviews of Educational Apps

Study	Type of review	Findings
Hirsh-Pasek et al. (2015)	Literature review	Conceptual framework for defining high-quality activities on educational apps
Haßler et al. (2016)	Literature review	Randomized trials and longitudinal studies needed to strengthen evidence base
Jamshidifarsani et al. (2019)	Analytical review	Content analysis of instructional mechanisms in problems
McTigue et al. (2020)	Critical review and meta-analysis	Meta-analysis of game-based literacy app, GraphoGame, found no significant main effect on word reading
Notari et al. (2016)	Literature review	Taxonomy to define educational apps
Papadakis et al. (2018)	Content analysis	Apps available through Google play promote rote learning rather than deeper conceptual understanding
Griffith et al. (2020)	Narrative synthesis	Apps for preschool-aged children confer an advantage in some domains (math)

that provide additional practice for students to master basic skills. Similar to tutoring interventions, apps may provide young children with more time on task and supplemental supports to master basic literacy and math skills (Nickow et al., 2020). Third, app designers are increasingly incorporating principles of gamified learning (Chou, 2016) such as learning goals, interactive activities, scaffolding, and rewards. Recent meta-analyses of digital games and gamified learning have shown medium-sized impacts on student learning and motivation outcomes (Clark et al., 2016; Sailer & Homner, 2020; Wouters et al., 2013). Importantly, educational apps may afford opportunities for developers to personalize learning by helping children and adults select appropriately leveled activities that support co-engagement with math content (Berkowitz et al., 2015). Although touchscreens, mobile devices, and computers that run educational apps are a ubiquitous feature of children's homes and classrooms (Clarke, 2014; Rideout & Robb, 2020), no meta-analysis to date has examined the potential effects, noneffects, or adverse effects of educational apps on children's academic skills or explored the sources of treatment heterogeneity.

What Are the Main Effects of Educational Apps on Literacy and Math Skills?

This meta-analytic review was motivated by two aims. Our first aim was to examine whether and to what extent educational apps produced positive and consistent main effects on preschool to Grade 3 students' literacy and math outcomes. We hypothesized that educational apps would improve both literacy and math outcomes by providing targeted opportunities for children to practice and develop academic skills that supplement traditional instruction particularly in school and classroom contexts. This hypothesis was based on meta-analytic reviews of one-to-one tutoring and small-group instruction provided by teachers, parents, or volunteers (Lipsey et al., 2012; Nickow et al., 2020) that demonstrate small and medium-sized effect sizes in literacy (ES = 0.35) and math (ES = 0.38).

What Study Characteristics Moderate the Effectiveness of Educational Apps?

Our second aim was to examine whether the effects of educational apps were moderated by methodological, participant, and intervention characteristics. Like other one-to-one tutoring and small-group interventions in the preschool and early elementary grades (Dietrichson et al., 2017), educational apps also vary along numerous methodological, participant, and intervention characteristics. Importantly, the average effect from a meta-analysis may conceal variability in treatment effects across studies. In particular, we explored the role of moderators that have been well known to explain variation in effect sizes in educational and behavioral intervention research, including the type of outcome, type of control condition, participants' grade level, and intervention dosage (Lipsey et al., 2012; Lipsey & Wilson, 1993). In addition, we examined the moderating role of intervention characteristics, particularly the quality of app activities and the type of skills they target (Hirsh-Pasek et al., 2015; McCormick et al., 2020).

Type of Assessment Outcome Measure and Control Group Activities. Prior research suggests that the type of outcome measure and control group activities moderate intervention impacts. In intervention studies involving preschool to Grade 3 children, average treatment effects are usually larger on researcher-developed measures that are closely tied to practice activities than standardized achievement tests (Lipsey et al. 2012; Paris, 2005). In many ways, improvement on a standardized outcome measure provides an index of far transfer (Barnett & Ceci, 2002; National Academies of Sciences, Engineering, and Medicine, 2018), highlighting whether students have mastered a broad domain of transferable knowledge that is not overly aligned with intervention activities (Lipsey et al., 2012; R. Wolf et al., 2020).

In addition to the type of assessment outcome, primary studies often find that the nature of the counterfactual may influence the magnitude of mean effects. That is, when studies compare educational apps to an active placebo group

rather than a passive group that is untreated, the magnitude of the treatment contrast in student outcomes may be attenuated (Griffith et al., 2020; Xie et al., 2018). For example, intervention studies of educational apps in math can include active placebo group activities where children in the control condition receive a literacy app (e.g., Berkowitz et al., 2015), or vice versa (e.g., Neuman, 2015). In an active placebo condition, there is a more rigorous test of the content of the app activities since both treatment and control students are completing educational activities utilizing the same medium.

Participants' Grade. Next, we examined whether the effectiveness of educational apps depends on the grade level of participating students in light of correlational research that paints a mixed portrait of whether educational apps, in particular, and screen time, in general, can help or hurt young children's academic achievement. Past research has focused on highlighting the effects, non-effects, and potential adverse effects of screen time and app usage with young children and has typically focused on either preschool (e.g., Griffith et al., 2020) or K–12 students (e.g., Cheung & Slavin, 2012). To our knowledge, no studies have attempted to compare mean effects for preschool and school-aged children. For example, some large-scale correlational studies have suggested that excessive screen time may have unintended negative consequences on young children's language and literacy development, communication skills, and socioemotional and health outcomes (Hutton et al., 2020; Madigan et al., 2019). In other words, the quality of the activities that children participate in may matter as much as the amount of time using mobile or interactive technologies (American Academy of Pediatrics, 2016). Accordingly, some policymakers (World Health Organization, 2019) have recommended that caregivers of preschool-aged children (3–4 years old) provide no more than 1 hour of sedentary screen time and the use of high-quality apps should ideally promote shared use and high-quality language interactions.

On the other hand, some scholars have argued that young children can thrive in a digital world where screen time and apps are a normal feature of daily life in school and home (Shapiro, 2018). A synthesis that focused on the effects of touchscreen devices found more promising evidence that young children could benefit from touchscreen devices but did not attempt to isolate the particular effects of educational apps on student achievement outcomes (Xie et al., 2018). A question that has yet to be explored is whether the effectiveness of educational apps depends on the participants' grade level. Therefore, we examined whether educational apps would be more or less effective for children in preschool versus kindergarten to Grade 3.

Intervention Dosage. An important malleable factor under the control of app designers and researchers is the amount of time that children are expected to work on an educational app. Existing research provides mixed findings on the relationship between intervention dosage and student outcomes. For example, meta-analytic evidence from tutoring studies involving one-to-one and small-group instruction has revealed limited differences in mean effects based on varying measures of intervention dosage such as the number of days per week or the total number of weeks that programs are offered to students (Nickow et al., 2020). The relationship between app usage on mobile and interactive technologies and student outcomes remains suggestive because findings are largely informed by nonexperimental research. For example, some correlational evidence indicated that more screen time may predict lower student achievement scores for both younger and older students (Hutton et al., 2020, World Health Organization, 2019), but correlational and survey research does not provide direct evidence on the causal effects of time spent using educational apps on student learning (Rideout, 2017; Kris, 2015; Livingstone, 2016).

Quality of App Activities and the Skills They Target. Importantly, there is growing evidence that educational apps must include high-quality activities that rest on research-based principles for improving learning more generally. In particular, educational apps should foster (a) active, engaged, and meaningful learning, supported by high-quality social interactions and clear learning goals (Hirsh-Pasek et al., 2015), and (b) deliberate practice that is focused, is active, includes regular feedback, and interleaves varied activities across different contexts (Bjork, 1994; Pashler et al., 2007).

Notably, researchers and developers have begun to develop apps that incorporate principles on how people learn and tested their efficacy in real-world settings. Berkowitz et al. (2015) conducted a randomized controlled trial (RCT) of the *Bedtime Learning Together* math app, which fosters co-engagement between children and parents around daily math word problems and led to improvements in unconstrained math skills. Other educational apps such as *Learn With Homer* are designed to improve constrained literacy schools by providing children games to support phonological skills in the context of structured lessons with the support of adults who monitored implementation fidelity (Neuman, 2015). Both of these illustrative examples of high-quality apps suggest the varied skills that are targeted by educational apps. Accordingly, we examined whether educational apps in literacy and math had larger effects on constrained rather than unconstrained skills (Lipsey et al., 2012; Lipsey et al., 2018; McCormick et al., 2020).

Method

Selection Criteria and Literature Search Procedures

The studies included in our review met the following five selection criteria. Each included study had to (a) evaluate the

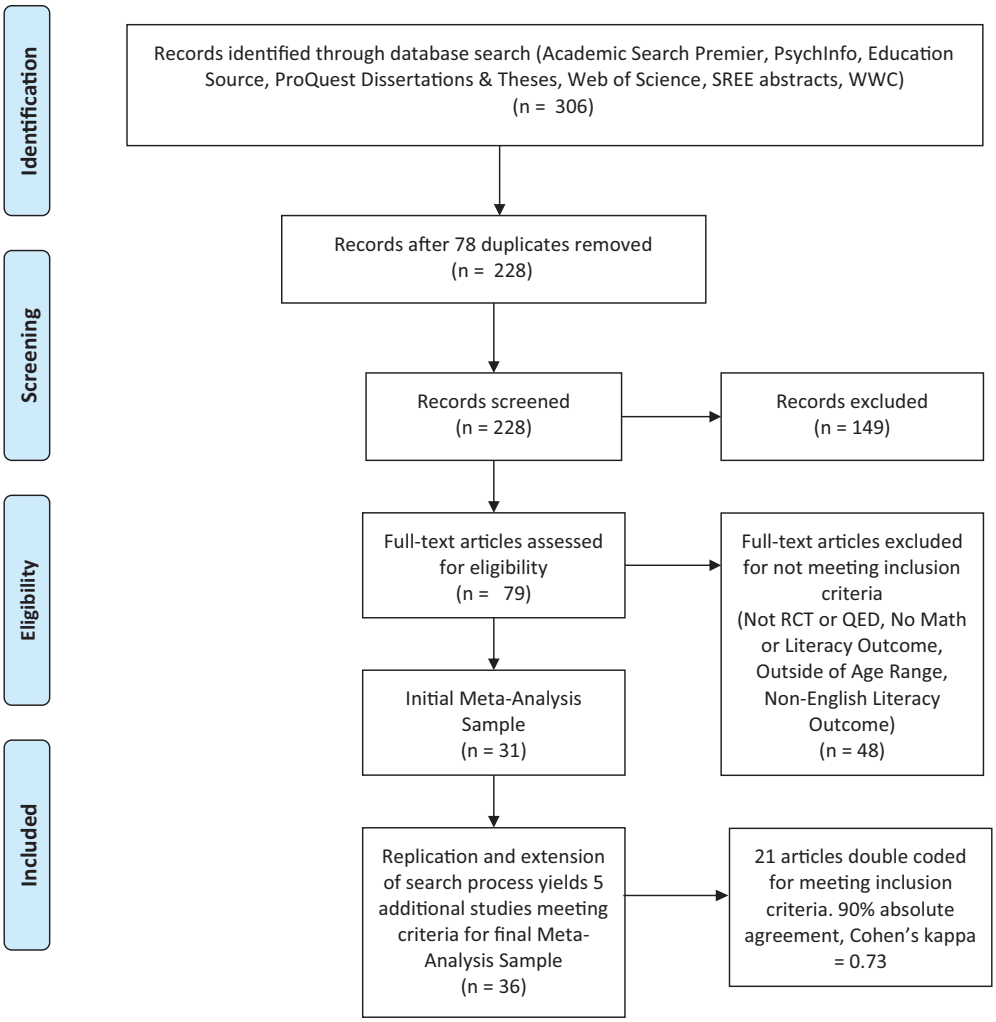

FIGURE 1. *Visual representation of the literature search and inclusion results.*
Note. WWC = What Works Clearinghouse; RCT = randomized controlled trial; QED = quasi-experimental design.

effects of an interactive educational app, (b) include an outcome measure of math or English language literacy skills, (c) provide sufficient empirical information to calculate an effect size, (d) include students from preschool to Grade 3 (approximately ages 3–9), and (e) use an experimental or quasi-experimental design to compare the postprogram performance of treatment students to control students who participated in either an active placebo or passive control group activity. We excluded studies using single-group pre-posttest designs because they fail to protect against most threats to internal validity (Shadish et al., 2002).

To identify primary studies, we searched (a) electronic databases and targeted internet sites, (b) reference lists of previous research syntheses, and (c) ancestral searches based on reference lists of included articles. Because the original iPhone was released in 2007, followed by Apple's App Store and Google Play in 2008, we limited our search to studies published in English from January 2008 to June 2020.

Electronic Databases

Figure 1 displays a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) screening flowchart describing our literature searching procedures. To identify published and unpublished studies, we searched electronic databases (Academic Search Premier, PsyInfo, Education Source, ProQuest Dissertations and Theses, Web of Science) and identified an initial sample of 306 studies. We also conducted searches of the gray literature by handsearching abstracts from annual meetings for the Society of Research on Educational Effectiveness and the What Works Clearinghouse's reviews of early literacy and math intervention studies. A full list of keywords for our searches is available in the online supplemental materials (Appendix 1). During the screening phase, we removed 78 duplicates, 149 studies that failed to meet inclusion criteria based on our review of the titles and abstracts, and 48 studies after we reviewed the full-text articles. An initial sample of 31

TABLE 2
Descriptive Characteristics of Included Studies (N = 36)

Publication characteristics (categorical)	*N*	%	*M* (*SD*), minimum/ maximum
Year of report			
2010	1	3	
2011	0	0	
2012	0	0	
2013	1	3	
2014	1	3	
2015	5	14	
2016	9	25	
2017	8	22	
2018	9	25	
2019	2	6	
Source of report			
Peer-reviewed journal article	24	67	
Book or chapter	0	0	
Dissertation	3	8	
MA thesis	0	0	
Private report	6	17	
Government report	0	0	
Conference paper	1	3	
Other	2	6	
Research design			
Randomized controlled trial (RCT)	33	92	
Quasi-experimental design	3	8	
Domain			
Literacy	12	33	
Math	24	67	
Type of skills targeted			
Constrained	22	61	
Unconstrained	14	39	
Assessment type			
Standardized test	14	39	
Researcher developed test	16	44	
Mix	6	17	
Type of control activities			
Active placebo control (yes)	10	28	
Active placebo control (no)	24	67	
Mix of active and passive control	2	6	
Grade level of participants			
Preschool	15	42	
Kindergarten	7	19	
1st grade	6	17	
2nd grade	3	8	
3rd grade	2	6	
Mix	2	6	
Unknown	1	3	
Intervention characteristics			
Frequency (*N* sessions)			32.1 (27.5), 1/120
Intensity (minutes per session)			21.1 (9.9), 10/50
Duration (length in days)			87.4 (68.3), 1/270
Quality of educational app (average)			2.4 (0.30), 1.8/2.8
Active			2.65 (0.48), 2/3
Engaging			2.79 (0.40), 2/3
Meaningful			2.32 (0.72), 1/3
Social			1.30 (0.52), 1/3
Learning Goal			2.93 (0.24), 2/3

included studies published from January 2008 to June 2019 was identified. We replicated this search process to update the review through June 2020 and found five additional studies that contributed to the final sample of 36 studies.

Procedures for Coding Studies

We developed a codebook to extract information from each of the 36 studies. The codebook was based on previous meta-analytic reviews of literacy intervention studies and prior research indicating the factors that would influence student outcomes (Durlak et al., 2011; Guo et al., 2020; J. S. Kim & Quinn, 2013; Lipsey et al., 2012; Marulis & Neuman, 2013). In particular, we coded for the content of educational app (math or literacy) and key methodological, participant, and intervention characteristics.

Table 2 indicates that over 90% of the studies were published in the past 5 years (2015–2020), suggesting this review is providing the most updated information on the effectiveness of educational apps. Most studies were published in peer-reviewed journal articles (67%), employed RCT designs (92%), and were closely split between preschool (42%) and K–3 samples (58%). In terms of discrete grade level, 42% were preschool, 19% in kindergarten, 17% in Grade 1, 8% in Grade 2, and 6% each in Grade 3 and mixed grades. There was also substantial variability in the mean quality of educational apps ($M = 2.40$, $SD = 0.48$), with a particularly low mean score for the social interaction indicator ($M = 1.30$, $SD = 0.52$). Table 3 provides additional details for each of the 36 studies.

Methodological Moderator Variables. Two critical methodological features that influence student outcomes are the type of outcome measure and counterfactual activities. First, because prior syntheses provide substantial evidence that mean effects would be larger for measures developed by study authors than standardized outcome measures (R. Wolf et al., 2020), we dichotomously coded for the type of outcome used in the study. Researcher-developed outcomes

were aligned with the intervention activities and measured more narrow domains of knowledge on specialized topics. In contrast, standardized outcomes were less aligned with the intervention and assessed broader domains of transferable knowledge (Kraft, 2019; Lipsey et al., 2012). Second, as described earlier, we coded for whether studies used an active placebo group—where control students completed activities on an app targeting a different domain—or a passive, "business-as-usual" control group. Approximately one half of the studies used standardized outcomes and one quarter used active placebo control groups.

Participant Moderator Variables. To compare mean effects by grade level, we coded for the grade level/age of participating students and created a dichotomous code indicating whether the sample was preschool- or school-aged (there were no studies that included both prekindergarten and older children).

Intervention Moderator Variables. To determine the moderating role of intervention features, we coded for three features. First, we coded for intervention dosage, or the amount of app usage. Based on prior research (Marulis & Neuman, 2013), we coded for (a) frequency, that is, the total number of sessions during the intervention; (b) intensity, that is, the length of each session in minutes; and (c) duration, that is, the length of the intervention from beginning to end. On average, studies included 32 sessions at about 21 minutes per session over the course of 87 days.

Second, we created an overall app quality score to assess the extent to which an educational app fostered learning that was active, engaging, meaningful, socially interactive, and had clear learning goals (Hirsh-Pasek et al., 2015). We downloaded each educational app in our review where possible and rated the following five criteria: (a) Do the activities promote active learning? (active), (b) Do the activities promote engaging learning? (engaging), (c) Do the activities promote meaningful learning? (meaningful), (d) Do the activities promote social interactions between children and adult caregivers? (social interaction), and (e) Do the activities have clear learning goals that foster educational aims? (learning goals). Each dimension was scored as low (e.g., app has no well-defined learning objective and is purely for entertainment), moderate (e.g., app has a vague literacy or math learning objective), or high (e.g., app has a clearly defined learning objective). When apps were not available for download, we used YouTube videos of app demonstrations or narrative descriptions of app functionality included in the research articles or supplementary online materials provided by app developers to assess the skills measured by the app. The mean quality score in our sample was 2.4. Details on the scoring rubric are included in the online supplemental materials (Appendix 2).

Third, we coded for the skills targeted and measured by primary researchers. Using previous coding systems (McCormick et al., 2020; Snow & Matthew, 2016), we coded both the skills targeted by the app and the skills that were measured by the outcomes used in the study. Constrained skills in both literacy and math can be improved by direct teaching, have a ceiling, and are mastered by most typically developing children. In contrast, unconstrained skills develop over time, require more varied experience, and are critical to higher order and more complex tasks like math problem solving and reading comprehension.

Publication Bias. One challenge of the vastly expanding market of apps and relevant research is that some high-quality studies that met our criteria and addressed our research questions may not be published through peer-reviewed academic channels. These alternative sources are known as "gray literature" (Marsolek et al., 2018) and can present issues for creating a truly systematic review of the literature. We therefore tested for publication bias and file drawer effects by (a) testing a moderator effect of a dichotomous published peer-reviewed article indicator, (b) using a trim and fill analysis (Duval & Tweedie, 2000), and (c) plotting a cumulative meta-analysis forest plot (Borenstein et al., 2009).

Coder Reliability. We created a codebook to collect information from each study and developed a procedure for estimating the reliability of the study codes. Two raters coded all moderator variables in our sample of 36 studies. Kappa coefficients adjust for chance agreement between raters and the mean kappa was $k = .94$ across coded study characteristics. All coding disagreements were resolved in follow-up meetings between coders.

Analytic Strategy

Calculation of Effect Sizes. To conduct a meta-analysis of continuous outcomes such as math and literacy achievement scores, we computed a standardized mean difference, or effect size. For each study, we computed Hedges's *g*—defined as the difference between the posttest means for the treatment and control group divided by the pooled standard deviation, with an adjustment for sample size. The effect sizes we analyzed come from our own calculations based on reported post-test means and standard deviations. Where these were not reported, we converted from reported test statistics to compute the effect size.

Robust Variance Estimation. First, we used robust variance estimation (RVE) to adjust standard errors to account for the correlation among effect sizes with studies. RVE allows syntheses to avoid a loss of information resulting from computing an aggregated, within-study average effect size. Following Tanner-Smith and Tipton (2014, p. 17), we

TABLE 3
Descriptive Characteristics of Each of the 36 Educational Apps

Author(s)	App name	Design: RCT or QED	Grade level	Context	*N*	Domain	Skill: C or UC	Quality score	ES	95% CI	Control
Alade et al. (2016)	Measuring With Murray	RCT	Preschool	School	60	Math	Unconstrained	2.6	0.658	[0.108, 1.208]	Active
Aunio & Mononen (2018)	Lola's World	RCT	Preschool	School	22	Math	Unconstrained	1.8	−0.041	[−0.938, 0.856]	Mix
D. L. Baker et al. (2017)	GraphoGame Spanish	RCT	1st	School	78	Literacy	Constrained	2.0	−0.191	[−0.644, 0.263]	Passive
Berkowitz et al. (2015)	Bedtime Math	RCT	1st	Home	586	Math	Unconstrained	2.6	0.100	[−0.062, 0.262]	Active
Cary et al. (2014)	KinderTek	RCT	K	School	94	Math	Constrained	2.4	0.207	[−0.200, 0.613]	Active
Cary et al. (2019)	KinderTek	RCT	K	School	473	Math	Constrained	2.4	−0.056	[−0.242, 0.130]	Passive
Cornu et al. (2017)	Unnamed Math App	RCT	K	School	125	Math	Constrained	2.2	0.142	[−0.212, 0.497]	Passive
Faizal (2016)	Conversation Gambits	RCT	Unknown	School	54	Literacy	Unconstrained	2.4	0.866	[0.307, 1.425]	Passive
Ferdig & Kosko (2017)	Zorbit	RCT	K	School	89	Math	Unconstrained	2.8	0.121	[−0.310, 0.551]	Passive
Fletcher (2015)	Unnamed math app	RCT	Mixed	School	85	Math	Unconstrained	2.2	0.342	[−0.089, 0.773]	Active
Gillis (2017)	Talking Shapes	RCT	Preschool	School	55	Literacy	Constrained	2.6	0.690	[0.139, 1.241]	Passive
Hallstedt et al. (2018)	Chasing Planets	RCT	2nd	School	281	Math	Constrained	2.4	0.225	[−0.012, 0.461]	Mix
Head (2016)	Various	RCT	2nd	School	77	Literacy	Unconstrained	2.6	−0.035	[−0.486, 0.417]	Passive
N. Kim et al. (2018)	123 Bakery	RCT	1st	School	46	Math	Unconstrained	2.4	0.392	[−0.197, 0.981]	Passive
Kosko & Ferdig (2016)	Zorbit	RCT	Preschool	Home	73	Math	Unconstrained	2.8	0.327	[−0.135, 0.788]	Passive
Kyle et al. (2013)	GG Rime; GG-Phoneme	RCT	1st	School	31	Literacy	Constrained	2.0	0.332	[−0.426, 1.090]	Passive
McManis & McManis (2016)	iSS	QED	Preschool	School	125	Literacy	Unconstrained	2.4	0.208	[−0.146, 0.563]	Passive
Neuman (2015)	Learn with Homer	RCT	Preschool	School	82	Literacy	Constrained	2.8	0.307	[−0.128, 0.743]	Active
Neumann (2018)	Various	RCT	Preschool	School	48	Literacy	Constrained	2.5	0.375	[−0.196, 0.947]	Passive
Outhwaite et al. (2018)	Maths 3–5, Maths 4–6	RCT	Preschool	School	389	Math	Unconstrained	2.6	0.014	[−0.196, 0.224]	Passive
Papadakis et al. (2018)	Unnamed math app	RCT	K	School	365	Math	Constrained	2.8	0.097	[−0.128, 0.321]	Passive
Patchan & Puranik (2016)	Writing Wizard	RCT	Preschool	School	46	Literacy	Constrained	1.8	0.806	[0.176, 1.436]	Passive
Patel et al. (2018)	GraphoLearn India	RCT	3rd	School	29	Literacy	Constrained	2.0	0.537	[−0.229, 1.302]	Active
Pitchford (2015)	Masamu	RCT	Mixed	School	42	Math	Constrained	2.6	0.468	[−0.151, 1.086]	Active
Presser et al. (2015)	Next Generation Preschool Math	RCT	Preschool	School	169	Math	Constrained	2.8	0.508	[0.201, 0.814]	Passive
Ramani et al. (2017)	The Great Race	RCT	K	School	54	Math	Constrained	2.4	0.073	[−0.464, 0.609]	Passive
Ramani et al. (2019)	The Great Race	RCT	K	School	96	Math	Constrained	2.4	0.058	[−0.339, 0.455]	Active
Schacter & Jo (2016)	Math Shelf	QED	Preschool	School	201	Math	Constrained	2.6	1.069	[0.704, 1.434]	Passive
Schacter & Jo (2017)	Math Shelf	RCT	Preschool	School	378	Math	Constrained	2.6	0.938	[0.721, 1.156]	Passive
Schacter et al. (2016)	Math Shelf	RCT	Preschool	School	86	Math	Constrained	2.6	0.634	[0.200, 1.068]	Active
Silander et al. (2016)	PEG+CAT	RCT	Preschool	Home	172	Math	Unconstrained	2.2	0.124	[−0.172, 0.42]	Passive
Szkudlarek & Brannon (2018)	123 Ninja	RCT	Preschool	School	105	Math	Unconstrained	2.4	−0.070	[−0.454, 0.314]	Active
Torgesen et al. (2010)	LIPS; RWT	RCT	1st	School	108	Literacy	Constrained	2.2	0.495	[0.096, 0.895]	Passive
Volk et al. (2017)	Clock, Map	QED	3rd	School	259	Math	Unconstrained	2.2	0.449	[0.202, 0.697]	Passive
Worth et al. (2018)	GG Rime	RCT	2nd	School	362	Literacy	Constrained	2.2	−0.048	[−0.254, 0.159]	Passive
van der Ven et al. (2017)	Unnamed math app	RCT	1st	School	102	Math	Constrained	2.2	0.349	[−0.043, 0.74]	Passive

Note. RCT = randomized controlled trial; QED = quasi-experimental design; ES = effect size; CI = confidence interval; LIPS = Lindamood Phoneme Sequencing; RWT = Read, Write & Type.

TABLE 4
Results of Estimating Unconditional Meta-Regression Model with RVE and Aggregated Mean Effects

Outcome	Estimation method	*k* Studies	*n* Effect sizes	Effect size	95% CI	*z*	Q_{total}	I^2	τ^2
Overall	RVE	36	285	0.31	[0.20, 0.42]	5.68**		0.83	0.15
Overall	Aggregated	36	36	0.30	[0.19, 0.41]	5.27**	121.22	0.71	0.07
Literacy	RVE	12	93	0.35	[0.13, 0.57]	3.46**		0.87	0.38
Literacy	Aggregated	12	12	0.31	[0.11, 0.51]	2.99**	25.05	0.56	0.06
Math	RVE	24	192	0.29	[0.16, 0.43]	4.50**		0.80	0.10
Math	Aggregated	24	24	0.29	[0.16, 0.43]	4.27**	95.6	0.76	0.08

Note. RVE = robust variance estimation; CI = confidence interval.
**$p < .01$.

applied RVE to our data set by computing weights for effect size *i* in study *j* based on the mean of within-study sampling variances for study *j*, the estimate of the between-studies variance component, the number of effect sizes within study *j*, and the estimated within-study correlation between all pairs of effect sizes (which we estimated to be .80). In addition, RVE methods allow for analyses of moderator variables that varied both between- and within studies. The within-study moderators used in our study (e.g., the type of outcome assessment, the type of skills assessed by the app) were centered around the variables' mean within each study to estimate the within-study effects.

Using Aggregated Effects to Supplement RVE Analyses. Because studies of educational apps vary along a number of dimensions and because we were interested in making inferences back to the population of studies from which our studies were sampled, we used a random effects model to pool the study-specific effect sizes and to generate an aggregated effect size (DerSimonian & Laird, 1986). The random effects model includes both a within-study weight (inverse of the study variance) and a between-study variance component. We made an a priori decision to employ a random effects model, because we expected that the dispersion of effect sizes would reflect true variance in mean effects.

In our data set, the most common dependency among effect sizes within studies involved correlated effects, which arises when multiple effect size estimates measure a single construct. Therefore, in addition to RVE analyses, we created an aggregated mean effect (i.e., a single average effect size for each study) to synthesize mean effects and to assess the robustness of results across two analytic methods. Using an aggregated mean effect size allowed us to maintain the assumption of statistical independence and to report heterogeneity statistics that are not available with RVE.

Measures of Heterogeneity. The meta-analysis of aggregated mean effects allows us to report the Q_{total} value, which tests the null hypothesis that mean effects shared a common effect (Borenstein et al., 2009). In addition, results based on both RVE and aggregated effects yield the I^2 value, which indicates the proportion of observed variance that represents true heterogeneity among studies along a 0 to 100% scale (Higgins et al., 2003), and the τ^2 estimate, which denotes the variance in mean effects. We used the *metan* package in Stata (StataCorp, 2019) and the meta package in R (R Core Team, 2020) to estimate the aggregated effects models, and the *robumeta* package in both Stata and R to estimate the RVE models. A replication toolkit including the data set and the Stata and R code is available from the authors upon request.

Sensitivity Analyses. To assess the sensitivity of our findings, we begin by reporting meta-analytic results from unconditional models that report combined impacts on overall achievement and separately for literacy and math. Next, we examine whether our results are replicated after controlling for whether the study was (a) published in a peer-reviewed journal, (b) used an RCT design, (c) used an active control group, and (d) used a standardized outcome assessment. We used the trim and fill method to assess the potential impact of missing, unpublished studies on mean effects (Duval & Tweedie, 2000).

Results

Main Effects of Educational Apps on Literacy and Math Skills

To address our first research aim, we used RVE and aggregated random effects models to synthesize findings from 36 studies and 285 effect sizes. As shown in Table 4, the RVE yielded a positive Hedges's *g* of +0.31 (95% confidence interval [CI] [0.20, 0.42]) and similar Hedges's *g* of +0.35 in literacy (95% CI [0.13, 0.57]) +0.29 in math (95% CI [0.16, 0.43]). Results were nearly identical for the random effects analysis of the aggregated effect sizes.

These mean effect sizes, however, mask substantial treatment effect heterogeneity. As shown by the results of the overall results of the aggregated effects model in Table 4, the Q_{total} statistic of 121.22 (degrees of freedom [*df*] = 35, $p < .001$) indicates that all studies do not share a common effect size. Furthermore, the I^2 statistic indicated that 71% of

TABLE 5
Results of Meta-Regression Model with RVE Controlling for Publication Status, Experimental Design, Control Group Activities, and Type of Assessment Outcome

Predictor	Estimate
Intercept	0.806**
	[0.231]
	3.484
Peer review	0.022
	[0.095]
	0.229
RCT	−0.194
	[0.221]
	−0.877
Active control	0.027
	[0.085]
	0.317
Standardized outcome	−0.419***
(study mean)	[0.086]
	−4.871
Math outcome	−0.183
	[0.094]
	−1.951
N effect sizes	285
k Studies	36
τ^2	0.13

Note. Cells are regression coefficients [*SE*s], and z values. RVE = robust variance estimation; RCT = randomized controlled trial.
*$p < .05$. **$p < .01$. ***$p < .001$.

the variance in observed effects reflected true between-study variability (rather than within-study sampling error). The estimated standard deviation in the true effects was almost as large as the mean effects overall ($\tau = .26$). The heterogeneity of the mean effect sizes across studies suggested that one or more study-level factors could moderate the impact of educational apps on student achievement outcomes.

Moderators of Educational App Effectiveness

Main Effects of Educational Apps Controlling for Between-Study Methodological Factors. The meta-regressions reported in Table 5 highlight potential methodological factors that moderated the impact of educational apps on student outcomes. Controlling for whether studies were published in peer-reviewed journals, used an RCT design, had active control groups, and targeted a literacy or math domain, the RVE meta-regression results indicate that studies using standardized outcomes produced impacts that were, on average, about 0.42 standard deviations ($p < .01$) lower than researcher developed outcomes. In addition, the results in table 5 suggest that no other methodological factors moderated effect sizes. Finally, because there were no differences in the impacts of educational apps that focused on literacy or math, we included both types of apps in all subsequent analyses.

Within- and Between-Study Moderators of Educational App Effectiveness. Table 6 displays a series of RVE meta-regression models that isolate the relationship between each respective participant and intervention moderator variable controlling for the type of outcome assessment. We controlled for whether outcomes were assessed with standardized measures in all models because it was a strong moderator of mean effects. Because standardized outcomes varied both within and between studies, we modeled separate within and between effects by including the study mean centered covariate along with the study mean value. Thus, all subsequent models include the controlled effects of participant and intervention characteristics. In Model 1, there was a statistically significant association between the mean effects and participants' grade, indicating that effects were 0.18 standard deviations higher, on average, in studies involving preschool-aged children than Kindergarten to Grade 3 children. Moving to Model 2, there was inconsistent evidence that intervention dosage measures were related to effect sizes. In particular, the meta-regressions indicated that $\log_2$ duration and $\log_2$ intensity measures did not predict outcomes once the assessment type was included in the model. Model 3 indicated that app quality ratings were not significantly associated with mean effect sizes, controlling for the type of assessment outcome. Furthermore, Model 4 indicated that the type of skill assessed by the apps was significantly associated with mean effects controlling for the type of assessment outcome.

After removing nonsignificant moderators of mean effects, we fit a final Model 5 that included participant grade, the type of outcome assessment (within and between studies), and the skills measured by educational apps. Importantly, the results of Model 5 indicated that educational apps produced mean effects that were approximately 0.17 *SD*s higher on constrained skills than on unconstrained skills (within or between studies), controlling for participant grade and the type of outcome assessment. To help interpret the findings from Model 5, we plotted predicted effect sizes based on the three statistically significant effect size moderators. Figure 2 shows the predicted mean differences for these three moderators holding the other covariates constant at their mean values. Predicted mean effects were larger for constrained skills ($g = 0.31$) than unconstrained skills ($g = 0.14$), for preschool samples ($g = 0.35$) than K–3 samples ($g = 0.17$), and for researcher-developed outcomes ($g = 0.43$) rather than standardized outcomes ($g = 0.17$).

Sensitivity Analyses for Publication Bias

We examined the effects of study design to explore the potential role of publication bias in two sensitivity analyses using the aggregated mean effect size per study as the unit of analysis. We assessed publication bias using the trim and fill analysis as shown in Figure 3 (Duval & Tweedie, 2000). There were six imputed study results that were missing on the left of the funnel plot that represented potentially unpublished studies with smaller mean effects. Imputing these six

TABLE 6
RVE Results With Intervention Characteristics as Moderators, Controlling for Type of Assessment Outcome

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5
Standardized test (within)	−0.142	−0.176	−0.141	−0.119	−0.113
	[0.202]	[0.264]	[0.201]	[0.202]	[0.202]
	−0.705	−0.667	−0.703	−0.591	−0.559
Standardized test (between)	−0.311***	−0.533***	−0.39***	−0.367***	−0.282***
	[0.09]	[0.138]	[0.094]	[0.097]	[0.089]
	−3.466	−3.854	−4.136	−3.787	−3.184
Log_2 duration		−0.012			
		[0.041]			
		−0.303			
Log_2 frequency		0.126*			
		[0.058]			
		2.163			
Log_2 intensity		0.012			
		[0.097]			
		0.127			
Average pillar score			0.213		
			[0.159]		
			1.336		
Preschool	0.178*				0.185**
	[0.091]				[0.086]
	1.953				2.151
Constrained outcome				0.137*	0.173**
				[0.076]	[0.07]
				1.808	2.479
Constant	0.382***	0.004	−0.01	0.42***	0.266***
	[0.069]	[0.409]	[0.39]	[0.082]	[0.08]
	5.522	0.01	−0.027	5.103	3.345

Note. Cells are regression coefficients [*SE*s], and z values. RVE = robust variance estimation.
*$p < .05$. **$p < .01$. ***$p < .001$.

mean effects yielded a mean effect size of $g = 0.20$ (observed + imputed effect sizes) compared to the mean effect size of $g = 0.30$ (observed effect size). Although there was a downward shift in the mean effect size, the substantive conclusion that educational apps have a positive mean effect remains unchanged. In addition, Figure 4 shows a cumulative forest plot, which displays mean effects based on studies with larger sample sizes and then shows whether and how much the mean effect size changes with the inclusion of smaller studies. The cumulative forest plot suggests that the mean effect size was stable as small-sample studies were added to the meta-analysis. In summary, these results increase our confidence that the findings are robust to potential publication bias.

Discussion

Educational apps are an increasingly ubiquitous feature of young children's lives at home and school, yet little is known about their effectiveness or the factors that diminish or enhance their impact on student achievement outcomes. Moreover, there is an urgent need to understand what works, for whom, and under what conditions as educators increasingly turn to easy-to-use technology interventions to support children's early literacy and math learning during the school closings triggered by the COVID-19 pandemic. To improve the rigor and relevance of the research base on educational apps, we undertook this meta-analysis of preschool to Grade 3 educational apps in math and literacy to advance two research goals. First, we examined the mean effects of 36 educational apps on preschool to Grade 3 children's math and literacy outcomes to quantify the extent to which apps improve student outcomes. Second, we examined the degree to which the effectiveness of educational apps was moderated by several methodological, participant, and intervention characteristics.

What Are the Main Effects of Educational Apps on Literacy and Math Skills?

In the domains of math and literacy, there was convergent evidence that educational apps improved student achievement outcomes relative to a counterfactual condition in which children participated in typical school instruction or

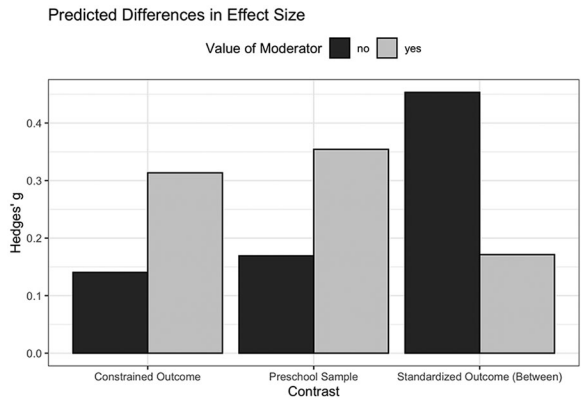

FIGURE 2. *Predicted differences in effect size.*
Note. Predicted differences in Hedges's *g* based on meta-regression Model 5 showing the magnitude of the three moderator effects: (1) skills targeted and measured by apps (constrained vs. unconstrained), (2) participants' grade (preschool vs. K–3), and (3) type of assessment outcome (standardized vs. researcher-developed outcome), controlling for other measured variables held constant at their means.

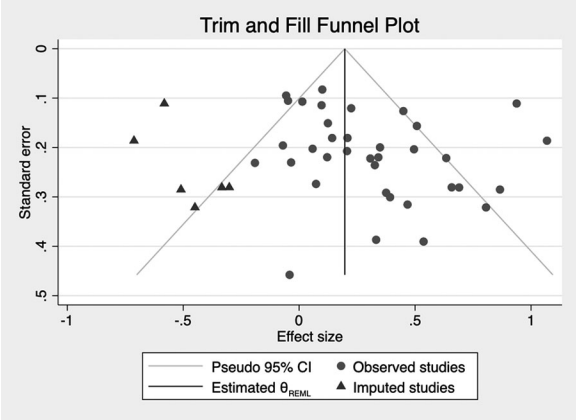

FIGURE 3. *Trim and fill plot.*
Note. Horizontal axis = effect size (Hedges's *g*); vertical axis = standard error; imputed studies are triangles. CI = confidence interval; REML = restricted maximum likelihood.

received a placebo control activity. Meta-analytic results based on RVE yielded medium-sized impacts in literacy (ES = 0.35) and math (ES = 0.29). The magnitude of these impacts is similar to the effects of tutoring interventions (ES = 0.37) and early elementary literacy interventions (ES = 0.39) based on recent meta-analyses of experimental and quasi-experimental intervention studies (Gersten et al., 2020; Nickow et al., 2020). Furthermore, the mean effects from our meta-analysis are consistent with recent meta-analyses of digital games and gamified learning interventions (Clark et al., 2016; Sailer & Homner, 2020; Wouters et al., 2013). These findings support theories that emphasize the affordances of educational apps in promoting active learning, deliberate practice, and gamified learning in one-to-one or small-group contexts to improve basic literacy and math skills (Griffith et al., 2020; Hirsh-Pasek et al., 2015).

Despite these promising findings, the 36 educational apps were unique in several ways. In particular, our meta-analysis included mostly high-quality apps that incorporated principles on how people learn, and the activities in turn were designed to promote learning that was active, engaging, meaningful, interactive, and focused on a clear learning goal (Hirsh-Pasek et al., 2015). For example, the educational apps in our review scored highly on all these dimensions ($M = 2.40/3.00$, $SD = 0.30$, range = 1.80–2.80) and were rigorously evaluated. Notably, 92% of the educational apps were used in school contexts. Although surveys of app use have tended to focus on the extent to which children use apps independently in home contexts (Radesky et al., 2020; Rideout, 2017), it is striking to note that the apps in our study were embedded into school contexts and routines. Finally, our meta-analytic strategy included a careful search of published and unpublished studies, studies with high internal validity, and analyses designed to rule out alternative explanations based on methodological artifacts or the choice of modeling strategies. In summary, our meta-analytic results are unlikely to generalize to apps that are neither based on principles from the science of learning nor subjected to rigorous experimental evaluations.

What Study Characteristics Moderate the Effectiveness of Educational Apps?

The potential pitfall of the current research base, however, is that the mean effects paint an overly simplistic and optimistic assessment of the value of educational apps. In many ways, the mean effect size overall, and separately for math and reading, masks true variance in mean effects. Given the dispersion in mean effects across studies, what were the key sources of treatment effect heterogeneity?

First, there was clear evidence that outcome measures matter. Whether a primary study used a researcher-designed or standardized outcome emerged as the most powerful moderator variable. For example, mean effects were nearly 0.28 *SD*s larger when primary studies used researcher-developed rather than standardized outcomes. The smaller magnitude of the mean effect size in studies using results on standardized outcome measures is in line with prior reviews of RCTs of elementary grade interventions (Elleman et al., 2009; Lipsey et al., 2012; Marulis & Neuman, 2013). A common explanation for this finding is that researcher-developed outcome measures assess more specialized domains of knowledge and are therefore easier to improve than standardized outcomes (Kraft, 2019; Lipsey et al., 2012; R. Wolf et al., 2020).

Second, meta-regressions that controlled for the type of assessment outcome revealed two additional moderators of app effectiveness. Controlling for assessment type and participant grade, measures of constrained skills produced mean effects that were 0.17 *SD*s higher, on average, than measures

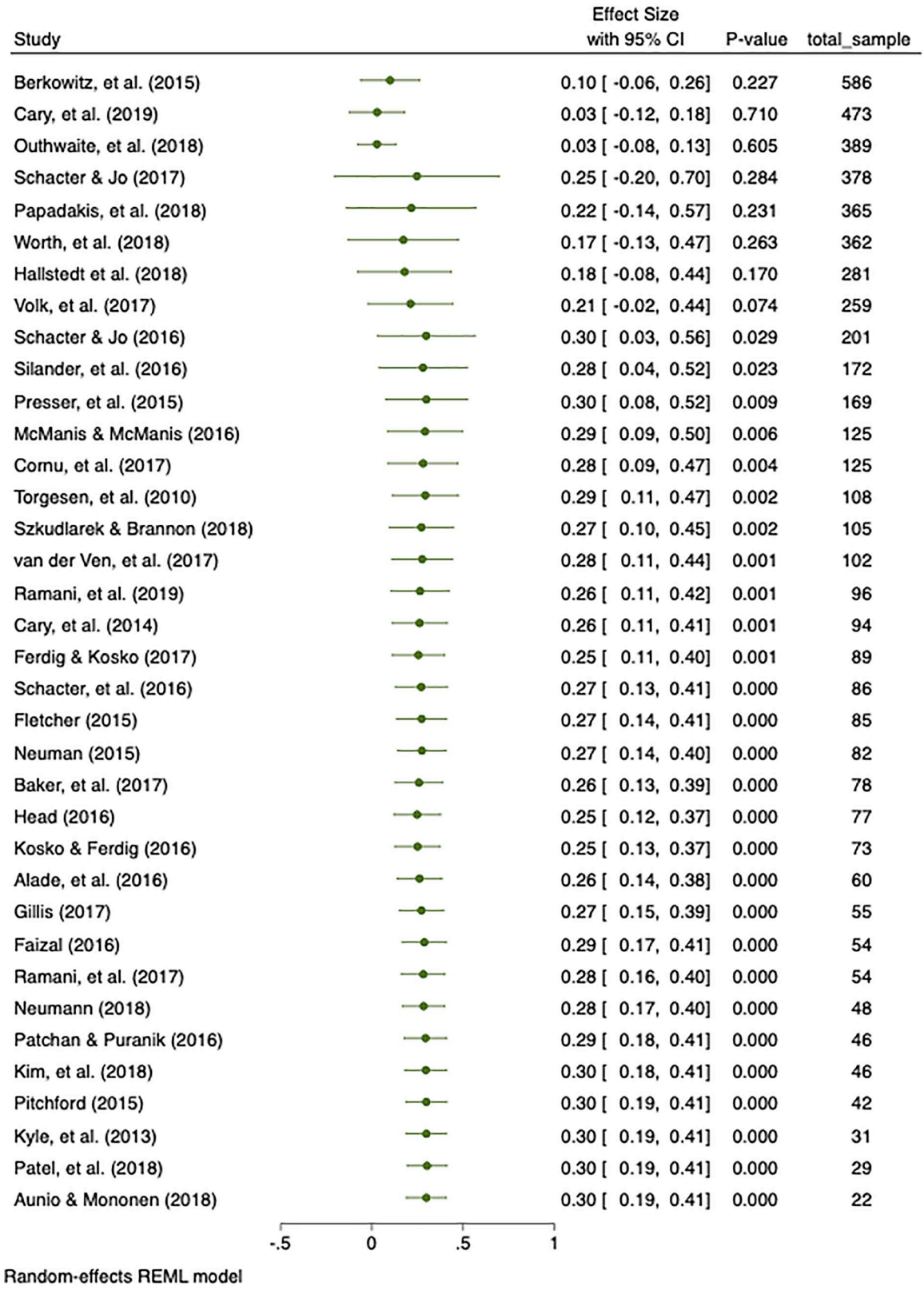

FIGURE 4. *Cumulative forest plot of aggregated mean effects from 36 studies.*
Note. CI = confidence interval; REML = restricted maximum likelihood.

of unconstrained skills, within and between studies. Given the short duration of the studies in our meta-analytic review, it is clear that many educational apps are designed primarily to improve constrained skills like number and letter recognition—that is, a small and fixed body of knowledge that is sensitive to short-term, targeted, and direct intervention efforts (Paris, 2005). To date, reviews of preschool educational apps suggest that most educational apps foster "simple drill and practice" activities designed primarily to improve constrained skills (Hirsh-Pasek et al., 2015; Papadakis et al., 2018). Consistent with other intervention research involving young children, there is growing evidence that educational interventions yield larger effects on literacy and math skills that are constrained to a small and fixed body of knowledge than unconstrained skills that require extensive practice and experience (Lipsey et al., 2012; Lipsey et al., 2018; Weiland & Yoshikawa, 2013). To a large extent, then, improving children's unconstrained skills—that is, mastery of broad domains like reading comprehension and math problem solving—will require longer and more intensive interventions (Brooks-Gunn et al., 2016).

Third, meta-regression results indicated that mean effects were larger in studies involving preschool-aged children than kindergarten to third-grade children. The magnitude of the preschool advantage over the K–3 grades is also consistent with recent meta-analytic results comparing effect sizes from RCTs of educational interventions with preschool- versus school-based children (Kraft, 2019; Lipsey et al., 2012). Although it is beyond the scope of this review to fully explain these findings, several hypotheses merit further scrutiny. One important hypothesis is that preschool-aged children may benefit from educational apps that foster joint attention and co-engagement in school contexts. Importantly, the majority of the studies involving preschool-aged children (87%, 13 of 15) were implemented in school contexts. This finding raises the question of whether and how co-engagement can enhance the effectiveness of educational apps in preschool center–based contexts (McTigue et al., 2020). During the preschool years, the essential condition for language development among young children is co-engagement and joint attention (Taylor, 2016), yet there is growing concern that mobile devices foster solo rather than co-use (through adult support) of educational activities that support literacy and math skills on interactive mobile technology (Bus et al., 2015; Radesky et al., 2014; M. Wolf, 2018). However, primary studies that directly compare the effectiveness of apps used in school and home contexts are needed given the increasing use of apps in children's non-school contexts (Rideout & Robb, 2020).

Finally, our meta-analytic results revealed no consistent association between intervention dosage and mean effect size across studies. These results are broadly consistent with reviews of recent intervention research finding no consistent association between measures of treatment dosage and student outcomes (Kraft et al., 2018; Lynch et al., 2019). How do we explain these findings? Most likely, our measure of quantity may reflect lower limits of time students spend on educational apps intended to run on mobile technology. That is, surveys of children's device usage report substantially higher levels of screen time than studies where the time on an app is tightly controlled for the duration of the intervention study. In other words, intervention research may not capture associations between the quantity of app usage and outcomes because the maximum is constrained by the end point of the study. Direct measures of children's app usage that leverage data from the mobile technology are needed to provide more accurate estimates of the time children spend using apps (Radesky et al., 2020; Roberts et al., 2016).

In addition to measuring relations between dosage and app effectiveness, we sought to measure the quality of the activities using principles from the learning sciences (Hirsh-Pasek et al., 2015; Pashler et al., 2007). These null findings, however, are inconclusive given the small number of studies in our review and the fact that most apps in our meta-analysis attempted to include research on how people learn. Accordingly, researchers should aim to pinpoint which aspects of activity, engagement, social interactions, and meaningfulness are most critical for supporting student learning outcomes. A strength of our review was the attempt to code five dimensions of app quality, which clearly indicated that few educational apps foster social interactions between children and their adult caregivers. Thus, one implication of our review is that quality depends on both the specific activities in an educational app and the rigor of the experimental design. For example, the *Bedtime Learning Together* math app (Berkowitz et al., 2015, average quality rating = 2.6 out of 3) and the *Learn With Homer* literacy app (Neuman, 2015, average quality rating = 2.8 out of 3) are both exemplar "high-quality" educational apps that were also tested using an RCT design with active placebo group. In addition, the Bedtime Learning Together math app is an example of a high-quality app that is designed to foster social interactions at home and has now been subjected to a long-term follow-up evaluation to determine whether short-term impacts fade out over time (Schaeffer et al., 2018). Importantly, educational apps that children and adults engage in together may have greater long-term effects than apps that children use alone. Building on this finding, researchers should continue to employ longitudinal and experimental designs to determine whether high-quality educational apps that foster co-engagement can produce enduring impacts on children's academic skills.

Limitations

Findings from this study highlight limitations that should inform future research. First, intervention studies have generated intent-to-treat estimates of the causal impact of offering children and parents an opportunity to use an educational app on a narrow set of outcomes. To our knowledge, no

study has attempted to connect measures of children's actual engagement in app activities to a wider set of outcomes. To improve the research base, more fine-grained measurement could inform practical recommendations about app use in particular and help shed light on potential adverse effects on children's social, emotional, and attentional outcomes. For example, descriptive research indicates that apps are simply one type of activity available to young children as they use smart phones and tablets to watch YouTube and play video games (Radesky et al., 2020). Direct measures of children's engagement (D'Mello et al., 2017)—that is, their actual time on task, their accuracy in completing digital activities, and their task orientations—are needed to connect the active ingredients in an educational app to a broader set of cognitive, behavioral, and motivational outcomes.

Second, there is a dearth of effectiveness research done at scale limiting the external validity of our findings. Current research on educational apps is lagging behind the push to scale-up easy-to-use remote learning interventions, which often include educational apps. For example, the U.S. Department of Education's Institute of Education Sciences has curated a website of over 100 apps that can be accessed for free (U.S. Department of Education, 2020), yet no evidence of effectiveness is provided for educators and parents. Notably, it is striking to find in our review that only six of the 36 educational apps in our study would meet ESSA (Every Student Succeeds Act) Tier I standards, requiring evidence from at least one well-executed RCT with more than 350 students.

Finally, we used stringent inclusion criteria to include only experimental and quasi-experimental evaluations of educational apps. Although our sensitivity analyses suggest that findings are robust to potential publication bias, there is a clear need to build on our review by casting a broader net that captures a wider set of unpublished studies. For example, we did not include evidence from the growing number of correlational studies that examine relationships between app use and student achievement. Leveraging school district administrative data involving 258 apps used by over 390,000 students, S. Baker and Gowda (2018) found that the average correlation between the amount of time students spent on educational apps was .01 in math and .00 in English language arts assessments. Although these results do not address selection bias, they underscore the need to combine descriptive, correlational, and experimental evidence in answering a broader set of questions and concerns facing decision-makers. In many ways, our study represents a first attempt to begin building a stronger foundation of evidence to understand whether and how educational apps deployed in real-world contexts can support student learning.

Conclusion

The purpose of this review was to synthesize results from experimental and quasi-experimental studies evaluating the impact of educational apps on children's math and literacy skills and to identify study characteristics that moderated mean effects. Although educational apps have positive effects on children's math and literacy skills, the effects are larger in studies involving preschool-aged children rather than kindergarten to Grade 3 children, studies using researcher-developed outcomes rather than standardized outcomes, and studies measuring constrained skills rather than unconstrained skills. Our findings suggest that the next generation of research on educational apps needs to improve both the internal and external validity of findings, evaluate effectiveness at much larger scale, use multiple outcome measures of student learning, and determine whether apps confer lasting benefits on a wider range of skills. Some efforts are currently under way to improve the quality of research (Molnar, 2020), but the marketplace for education apps remains "chaotic and unregulated" (Papadakis et al., 2018, p. 156). Although apps are increasingly advertised to educators and parents as a low-cost and scalable strategy for improving learning, apps are not uniformly high-quality and rarely evaluated rigorously by independent researchers.

More collaborative research across disciplinary silos is clearly needed to address these research gaps. In particular, we encourage learning scientists to incorporate principles on how people learn into the design of high-quality activities in literacy and math apps, developmental pediatricians and psychologists to examine how the quantity and quality of adult mediation in school and home contexts support learning, and intervention researchers to explore whether direct measures of engagement mediate the effects of educational apps on student outcomes (Cherner et al., 2014; Griffith et al., 2020; McTigue et al., 2020; Papadakis et al., 2018; Radesky et al., 2020). In short, we encourage the field to move beyond the broad question—do apps work—to the more targeted question: How and under what conditions do high-quality educational apps support children's early literacy and math skills? The limitations of this meta-analytic review highlight the collaborative research needed to shed light on this question.

ORCID iD

James Kim https://orcid.org/0000-0002-6415-5496

References

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

*Aladé, F., Lauricella, A. R., Beaudoin-Ryan, L., & Wartella, E. (2016). Measuring With Murray: Touchscreen technology and preschoolers' STEM learning. *Computers in Human Behavior*, *62*, 433–441.https://doi.org/10.1016/j.chb.2016.03.080

American Academy of Pediatrics. (2016). Media and Young Minds, Council on Communications and Media. *Pediatrics*, *138*(5), e20162591. https://doi.org/10.1542/peds.2016-2591

*Aunio, P., & Mononen, R. (2018). The effects of educational computer game on low-performing children's early numeracy

skills: An intervention study in a preschool setting. *European Journal of Special Needs Education*, *33*(5), 677–691. https://doi.org/10.1080/08856257.2017.1412640

*Baker, D. L., Basaraba, D. L., Smolkowski, K., Conry, J., Hautala, J., Richardson, U., English, S., & Cole, R. (2017). Exploring the cross-linguistic transfer of reading skills in Spanish to English in the context of a computer adaptive reading intervention. *Bilingual Research Journal*, *40*(2), 222–239. https://doi.org/10.1080/15235882.2017.1309719

Baker, S., & Gowda, S. (2018). *The 2018 Technology & Learning Insights Report: Towards understanding app effectiveness and cost*. BrightBytes.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

*Berkowitz, T., Schaeffer, M. W., Maloney, E. A., Peterson, L., Gregor, C., Levine, S. C., & Beilock, S. L. (2015). Math at home adds up to achievement in school. *Science*, *350*(6257), 196–198. https://doi.org/10.1126/science.aac7427

Bjork, R. A. (1994). Institutional impediments to effective training. In D. Druckman, & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 295–306). National Academies Press.

Borenstein, M., Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Brooks-Gunn, J., Markman-Pithers, L., & Rouse, C. E. (2016). Starting early: Introducing the issue. *The Future of Children*, *26*(2), 3–19. https://doi.org/10.1353/foc.2016.0009

Bus, A. G., Takacs, Z. K., & Kegel, C. A. (2015). Affordances and limitations of electronic storybooks for young children's emergent literacy. *Developmental Review*, *35*, 79–97. https://doi.org/10.1016/j.dr.2014.12.004

*Cary, M., Kennedy, P., Crowley, R., Shanley, L., & Clarke, B. (2019). *KinderTEK iPad Math Program: Results from Cohort 1*. Center on Teaching and Learning, University of Oregon.

*Cary, M., Shanley, L., Clarke, B., & Sota, M. (2014). *Evaluating the KinderTEK iPad App's individualized and adaptive math instruction*. Center on Teaching and Learning, University of Oregon.

Cherner, T., Dix, J., & Lee, C. (2014). Cleaning up that mess: A framework for classifying educational apps. *Contemporary Issues in Technology and Teacher Education*, *14*(2), 158–193.

Cheung, A. C., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review*, *7*(3), 198–215. https://doi.org/10.1016/j.edurev.2012.05.002

Chou, Y. (2016). *Actionable gamification: Beyond points, badges, and leaderboards*. Octaylysis Media.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Clarke, B. (2014). *The use of tablets in UK Schools: Stage 4*. Family, Kids and Youth Research Group. http://www.kidsandyouth.com/pdf/FK%26Y%20T4S%20Stage%204%20The%20Use%20of%20Tablets%20in%20UK%20Schools%20Sept%2014.pdf

*Cornu, V., Schiltz, C., Pazouki, T., & Martin, R. (2017). Training early visuo-spatial abilities: A controlled classroom-based intervention study. *Applied Developmental Science*, *23*(1), 1–21.https://doi.org/10.1080/10888691.2016.1276835

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A. M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*(2), 243–282. https://doi.org/10.3102/0034654316687036

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*(2), 104–123. https://doi.org/10.1080/00461520.2017.1281747

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1–44. https://doi.org/10.1080/19345740802539200

*Faizal, M. A. (2016, September). The effects of conversation-gambits visual-novel game on students' English achievement and motivation. In *2016 International Electronics Symposium (IES)* (pp. 481–486). IEEE.

*Ferdig, R., & Kosko, K. (2017). *Interactive learning games can significantly improve early math learning and opportunities for cross-grade growth* [White paper]. https://www.zorbitsmath.com/wp-content/uploads/2018/02/Zorbit-Kindergarten-White-Paper.pdf

*Fletcher, N. (2015). *Development and evaluation of a computer program to teach symmetry to young children* [Doctoral dissertation, Columbia University]. https://academiccommons.columbia.edu/doi/10.7916/D8RF5T4Z

Flewitt, R., Messer, D., & Kucirkova, N. (2015). New directions for early literacy in a digital age: The iPad. *Journal of Early Childhood Literacy*, *15*(3), 289–310. https://doi.org/10.1177/1468798414533560

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020) Meta-analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness*, *13*(2), 401–427. https://doi.org/10.1080/19345747.2019.1689591

*Gillis, M. (2017). *Talking shapes research*. Literacy How. https://www.talkingfingers.com/talking-shapes/

Griffith, S. F., Hagan, M. B., Heymann, P., Heflin, B. H., & Bagner, D. M. (2020). Apps as learning tools: A systematic review. *Pediatrics*, *145*(1), e20191579. https://doi.org/10.1542/peds.2019-1579

Guernsey, L., Levine, M., Chiong, C., & Severns, M. (2012). *Pioneering literacy in the digital wild west: Empowering parents and educators*. Campaign for Grade-Level Reading.

Guo, D., Zhang, S., Wright, K. L., & McTigue, E. M. (2020). Do you get the picture? A meta-analysis of the effect of graphics on reading comprehension. *AERA Open*, *6*(1), 2332858420901696. https://doi.org/10.1177/2332858420901696

Haßler, B., Major, L., & Hennessy, S. (2016). Tablet use in schools: A critical review of the evidence for learning outcomes. *Journal of Computer Assisted Learning*, *32*(2), 139–156. https://doi.org/10.1111/jcal.12123

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., & Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education*, *102*, 202–223. https://doi.org/10.1016/j.compedu.2016.09.001

*Hallstedt, M., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology*, *110*(8), 1127–1148. https://doi.org/10.1037/edu0000264

Head, T. S. (2016). *Supporting literacy with iPads: A pilot study in second-grade classrooms* [Doctoral dissertation]. ProQuest Dissertations and Theses Global.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps lessons from the science of learning. *Psychological Science in the Public Interest*, *16*(1), 3–34. https://doi.org/10.1177/1529100615569721

Hutton, J. S., Dudley, J., Horowitz-Kraus, T., DeWitt, T., & Holland, S. K. (2020). Associations between screen-based media use and brain white matter integrity in preschool-aged children. *JAMA Pediatrics*, *174*(1), e193869–e193869. https://doi.org/10.1001/jamapediatrics.2019.3869

Jamshidifarsani, H., Garbaya, S., Lim, T., Blazevic, P., & Ritchie, J. M. (2019). Technology-based reading intervention programs for elementary grades: An analytical review. *Computers & Education*, *128*, 427–451. https://doi.org/10.1016/j.compedu.2018.10.003

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*(3), 386–431. https://doi.org/10.3102/0034654313483906

*Kim, N., Jang, S., & Cho, S. (2018). Testing the efficacy of training basic numerical cognition and transfer effects to improvement in children's math ability. *Frontiers in Psychology*, *9*, Article 1775. https://doi.org/10.3389/fpsyg.2018.01775

*Kosko, K., & Ferdig, R. (2016). Effects of a tablet-based mathematics application for pre-school children. *Journal of Computers in Mathematics and Science Teaching*, *35*(1), 61–79.

Kraft, M. (2019). *Interpreting effect sizes of education interventions* (EdWorkingPaper No.19-10). http://www.edworkingpapers.com/ai19-10

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kris, D. F. (2015). *How to provide kids with screen time that supports learning*. https://ww2.kqed.org/mindshift/2015/11/11/how-to-provide-kids-with-screen-time-that-supports-learning/

Kucirkova, N. (2014). iPads in early education: Separating assumptions and evidence. *Frontiers in Psychology*, *5*, Article 715. https://doi.org/10.3389/fpsyg.2014.00715

*Kyle, F. E., Kujala, J., Richardson, U., Lyytinen, H., & Goswami, U. (2013). Assessing the effectiveness of two theoretically motivated computer-assisted reading interventions in the United Kingdom: GG Rime and GG Phoneme. *Reading Research Quarterly*, *48*(1), 61–76. https://doi.org/10.1002/rrq.038

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee prekindergarten program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly*, *45*, 155–176. https://doi.org/10.1016/j.ecresq.2018.03

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable Forms*. National Center for Special Education Research.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *The American Psychologist*, *48*(12), 1181–1209. https://doi.org/10.1037/0003-066X.48.12.1181

Livingstone, S. (2016). *What are pre-schoolers doing with tablets and is it good for them?* https://goo.gl/tyo9y2

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293. https://doi.org/10.3102/0162373719849044

Madigan, S., Browne, D., Racine, N., Mori, C., & Tough, S. (2019). Associations between screen time and children's performance on a developmental screening test. *JAMA Pediatrics*, *173*(3), 244–250. https://doi.org/10.1001/jamapediatrics.2018.5056

Marsolek, W., Cooper, K., Farrell, S., & Kelly, J. (2018). The types, frequencies, and findability of disciplinary grey literature within prominent subject databases and academic institutional repositories. *Journal of Librarianship and Scholarly Communication*, *6*(1), eP2200. https://doi.org/10.7710/2162-3309.2200

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, *6*(3), 223–262. https://doi.org/10.1080/19345747.2012.755591

McCormick, M. P., Weissman, A. K., Weiland, C., Hsueh, J., Sachs, J., & Snow, C. (2020). Time well spent: Home learning activities and gains in children's academic skills in the prekindergarten Year. *Developmental Psychology*, *56*(4), 710–726. https://doi.org/10.1037/dev0000891

*McManis, M. H., & McManis, L. D. (2016). Using a touch-based, computer-assisted learning system to promote literacy and math skills for low-income preschoolers. *Journal of Information Technology Education*, *15*, 409–429. https://doi.org/10.28945/3550

McTigue, E. M., Solheim, O. J., Zimmer, W. K., & Uppstad, P. H. (2020). Critically reviewing GraphoGame across the world: Recommendations and cautions for research and

implementation of computer-assisted instruction for word-reading acquisition. *Reading Research Quarterly*, *55*(1), 45–73. https://doi.org/10.1002/rrq.256

Molnar, M. (2020). *Got a research base? Ed-Tech certification offers companies a chance to prove it* (EdWeek Market Brief 6). https://marketbrief.edweek.org/marketplace-k-12/digital-promises-ed-tech-product-certification-centers-research-based-design/

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.

National Research Council 2015. *Transforming the workforce for children birth through age 8: A unifying foundation*. National Academies Press. https://doi.org/10.17226/19401.

*Neuman, S. B. (2015). *Closing the app gap: Improving children's phonological skills*. https://learnwithhomer.com/Closing-the-App-Gap.pdf

*Neumann, M. M. (2018). Using tablets and apps to enhance emergent literacy skills in young children. *Early Childhood Research Quarterly*, *42*, 239–246. https://doi.org/10.1016/j.ecresq.2017.10.006

Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. (EdWorkingPaper No. 20-267). https://doi.org/10.26300/eh0c-pc52

Notari, M. P., Hielscher, M., & King, M. (2016). Educational apps ontology. In D. Churchill, J. Lu, T. K. F. Chui, & B. Fox (Eds.), *Mobile learning design* (pp. 83–96). Springer. https://doi.org/10.1007/978-981-10-0027-0_5

*Outhwaite, L. A., Faulder, M., Gulliford, A., & Pitchford, N. J. (2018). Raising early achievement in math with interactive apps: A randomized control trial. *Journal of Educational Psychology*, *111*(2), 284–298. https://doi.org/10.1037/edu0000286

*Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2018). The effectiveness of computer and tablet assisted intervention in early childhood students' understanding of numbers. An empirical study conducted in Greece. *Education and Information Technologies*, *23*(5), 1849–1871. https://doi.org/10.1007/s10639-018-9693-7

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning: IES practice guide* (NCER 2007-2004). National Center for Education Research.

*Patchan, M. M., & Puranik, C. S. (2016). Using tablet computers to teach preschool children to write letters: Exploring the impact of extrinsic and intrinsic feedback. *Computers & Education*, *102*, 128–137. https://doi.org/10.1016/j.compedu.2016.07.007

*Patel, P., Torppa, M., Aro, M., Richardson, U., & Lyytinen, H. (2018). GraphoLearn India: The effectiveness of a computer-assisted reading intervention in supporting struggling readers of English. *Frontiers in Psychology*, *9*, Article 1045. https://doi.org/10.3389/fpsyg.2018.01045

Pearson, P. D., Palincsar, A. S., Biancarosa, G., & Berman, A. I. (Eds.). (2020). *Reaping the rewards of the Reading for Understanding Initiative*. National Academy of Education.

Pendlebury, T. (2018). *All the 2018 education apps Apple announced*. CNET. https://www.cnet.com/pictures/all-the-2018-education-apps-apple-announced/

*Pitchford, N. J. (2015). Development of early mathematical skills with a tablet intervention: A randomized control trial in Malawi. *Frontiers in Psychology*, *6*, Article 485. https://doi.org/10.3389/fpsyg.2015.00485

*Presser, A. L., Vahey, P., & Dominguez, X. (2015). *Improving mathematics learning by integrating curricular activities with innovative and developmentally appropriate digital apps: Findings from the next generation preschool math evaluation*. Society for Research on Educational Effectiveness.

Radesky, J. S., Kistin, C. J., Zuckerman, B., Nitzberg, K., Gross, J., Kaplan-Sanoff, M., Augustyn, M., & Silverstein, M. (2014). Patterns of mobile device use by caregivers and children during meals in fast food restaurants. *Pediatrics*, *133*(4), e843–e849. https://doi.org/10.1542/peds.2013-3703

Radesky, J. S., Weeks, H. M., Ball, R., Schaller, A., Yeo, S., Durnez, J., Tamayo-Rios, M., Epstein, M., Kirkorian, H., Coyne, S., & Barr, R. (2020). Young children's use of smartphones and tablets. *Pediatrics*, *146*(1), e20193518. https://doi.org/10.1542/peds.2019-3518

*Ramani, G. B., Daubert, E. N., Lin, G. C., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and working memory games on kindergartners' numerical knowledge. *Developmental Science*, *23*(4), e12908. https://doi.org/10.1111/desc.12908.

*Ramani, G. B., Jaeggi, S. M., Daubert, E. N., & Buschkuehl, M. (2017). Domain-specific and domain-general training to improve kindergarten children's mathematics. *Journal of Numerical Cognition*, *3*(2), 468–495. https://doi.org/10.5964/jnc.v3i2.31

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/

Richards, J., & Stebbins, L. (2014). *2014 U.S. education technology industry market: PreK-12*. Software & Information Industry Association.

Rideout, V. (2017). *The Common Sense census: Media use by kids age zero to eight*. Common Sense Media.

Rideout, V., & Robb, M. B. (2020). *The Common Sense census: Media use by kids age zero to eight, 2020*. Common Sense Media

Roberts, J. D., Chung, G. K., & Parks, C. B. (2016). Supporting children's progress through the PBS KIDS learning analytics platform. *Journal of Children and Media*, *10*(2), 257–266. https://doi.org/10.1080/17482798.2016.1140489

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*, 77–112. https://doi.org/10.1007/s10648-019-09498-w

*Schacter, J., & Jo, B. (2016). Improving low-income preschoolers' mathematics achievement with Math Shelf, a preschool tablet computer curriculum. *Computers in Human Behavior*, *55*(Pt. A), 223–229. https://doi.org/10.1016/j.chb.2015.09.013

*Schacter, J., & Jo, B. (2017). Improving preschoolers' mathematics achievement with tablets: A randomized controlled trial. *Mathematics Education Research Journal*, *29*(3), 313–327. https://doi.org/10.1007/s13394-017-0203-9

*Schacter, J., Shih, J., Allen, C. M., DeVaul, L., Adkins, A. B., Ito, T., & Jo, B. (2016). Math Shelf: A randomized trial of a prekindergarten tablet number sense curriculum. *Early Education and Development*, *27*(1), 74–88. https://doi.org/10.1080/10409289.2015.1057462

Schaeffer, M. W., Rozek, C. S., Berkowitz, T., Levine, S. C., & Beilock, S. L. (2018). Disassociating the relation between parents' math anxiety and children's math achievement: Long-term effects of a math app intervention. *Journal of Experimental Psychology: General*, *147*(12), 1782–1790. https://doi.org/10.1037/xge0000490

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shapiro, J. (2018). *The new childhood: Raising kids to thrive in a digitally connected world*. Hachette UK.

*Silander, M., Moorthy, S., Dominguez, X., Hupert, N., Pasnik, S., & Llorente, C. (2016). *Using digital media at home to promote young children's mathematics learning: Results of a randomized controlled trial*. Society for Research on Educational Effectiveness.

Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children*, *26*(2), 57–74. https://doi.org/10.1353/foc.2016.0012

StataCorp. (2019). Stata statistical software: Release 16.

Szkudlarek, E., & Brannon, E. M. (2018). Approximate arithmetic training improves informal math performance in low achieving preschoolers. *Frontiers in Psychology*, *9*, Article 606. https://doi.org/10.3389/fpsyg.2018.00606

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations and a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30. https://doi.org/10.1002/jrsm.1091

Taylor, C. (2016). *The language animal*. Harvard University Press.

*Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Herron, J., & Lindamood, P. (2010). Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: Outcomes from two instructional approaches. *Annals of Dyslexia*, *60*(1), 40–56. https://doi.org/10.1007/s11881-009-0032-y

U.S. Department of Education. (2020). *The ED games expo "goes virtual" to support distance learning*. https://ies.ed.gov/blogs/research/post/the-ed-games-expo-goes-virtual-to-support-distance-learning-82-us-department-of-education-and-government-supported-learning-games-and-technologies-are-now-available-at-no-cost-until-the-end-of-the-school-year

*Van der Ven, F., Segers, E., Takashima, A., & Verhoeven, L. (2017). Effects of a tablet game intervention on simple addition and subtraction fluency in first graders. *Computers in Human Behavior*, *72*, 200–207. https://doi.org/10.1016/j.chb.2017.02.031

*Volk, M., Cotič, M., Zajc, M., & Starcic, A. I. (2017). Tablet-based cross-curricular maths vs. traditional maths classroom practice for higher-order learning outcomes. *Computers & Education*, *114*, 1–23. https://doi.org/10.1016/j.compedu.2017.06.004

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

Wexler, N. (2019). How classroom technology is holding students back. *MIT Technology Review*. https://www.technologyreview.com/s/614893/classroom-technology-holding-students-back-edtech-kids-education

Wolf, M. (2018). *Reader, come home: The reading brain in a digital world*. Harper.

Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness*, *13*(2), 428–447. https://doi.org/10.1080/19345747.2020.1726537

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122–154. https://doi.org/10.1002/pam.20310

World Health Organization. (2019). *Guidelines on physical activity, sedentary behaviour and sleep for children under 5 years of age: summary*. https://apps.who.int/iris/handle/10665/325147

*Worth, J., Nelson, J., Harland, J., Bernardinelli, D., & Styles, B. (2018). *GraphoGame Rime*. https://www.graphogame.com/index.html

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of educational psychology*, *105*(2), 249–265. https://doi.org/10.1037/a0031311

Xie, H., Peng, J., Qin, M., Huang, X., Tian, F., & Zhou, Z. (2018). Can touchscreen devices be used to facilitate young children's learning? A meta-analysis of touchscreen learning effect. *Frontiers in Psychology*, *9*, Article 2580. https://doi.org/10.3389/fpsyg.2018.02580

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *The Future of Children*, *26*(2), 21–35. https://doi.org/10.1353/foc.2016.0010

Authors

JAMES KIM is a professor of education at Harvard Graduate School of Education. His current research priority is to understand how building children's domain knowledge and reading engagement can foster long-term improvements in reading comprehension.

JOSHUA GILBERT is a senior researcher at the Harvard READS Lab and is a faculty member at New England Conservatory in the Music-in-Education Department. His research interests include arts learning and quantitative methods.

CHARLES GALE is a PhD candidate in education policy and program evaluation at Harvard Graduate School of Education. His research interests include education finance, early literacy, and early childhood education.

QUN YU is a PhD candidate in curriculum and instruction at Boston College. Her research interests include children's language and literacy development in the early years and the effectiveness of literacy interventions.