EdWorkingPaper No. 21-379

The Impact of Summer Learning Programs on Low-Income Children's Mathematics Achievement: A Meta-Analysis

Kathleen Lynch
University of Connecticut

Lily An
Harvard University

Zid Mancenido
Harvard University

We present results from a meta-analysis of 37 experimental and quasi-experimental studies of summer programs in mathematics for children in grades pre-K-12, examining what resources and characteristics relate to stronger student achievement, attainment, and social-emotional and behavioral outcomes. Compared to control group children, children who participated in summer programs that included mathematics lessons and activities enjoyed significant improvements in mathematics learning as well as social-behavioral outcomes. We find an average weighted impact estimate of +0.09 standard deviations on mathematics achievement outcomes. In a parallel meta-analysis, we found similar positive impacts of summer programs on socialemotional and behavioral outcomes, Programs conducted in both high- and lower-poverty settings saw similar positive impacts. The results highlight the potential for summer programs to strengthen children's mathematical ability and improve learning outcomes in both mixed-poverty and high-poverty settings.

VERSION: April 2021

Suggested citation: Lynch, Kathleen, Lily An, and Zid Mancenido. (2021). The Impact of Summer Learning Programs on Low-Income Children's Mathematics Achievement: A Meta-Analysis. (EdWorkingPaper: 21-379). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/da7r-4z83

The Impact of Summer Learning Programs on Low-Income Children's Mathematics Achievement:
A Meta-Analysis

Updated: April 2021

Abstract

We present results from a meta-analysis of 37 experimental and quasi-experimental studies of summer programs in mathematics for children in grades pre-K-12, examining what resources and characteristics relate to stronger student achievement, attainment, and social-emotional and behavioral outcomes. Compared to control group children, children who participated in summer programs that included mathematics lessons and activities reaped significant improvements in mathematics learning as well as social-behavioral outcomes. We find an average weighted impact estimate of +0.09 standard deviations on mathematics achievement outcomes. We also undertook a smaller, secondary analysis of the effect of summer programs on social-emotional and behavioral outcomes, and found similar positive effects. We find no difference in effects for programs conducted in high and mid-low poverty settings. The results point toward potential for summer programs to strengthen low-income children's mathematical proficiency outside of school time.

Suggested Citation:

Lynch, K., An, L., & Mancenido, Z. (2021). The Impact of Summer Learning Programs on Low-Income Children's Mathematics Achievement: A Meta-Analysis. Working Paper.

The Impact of Summer Learning Programs on Low-Income Children's Mathematics Achievement: A Meta-Analysis

Three out of four low-income children in the U.S. fail to meet standards for mathematical proficiency in the fourth grade, as compared with only 43% of middle-income children (NAEP, 2017). This early inequality in math achievement is problematic because mathematical knowledge is cumulative (Hiebert & Wearne, 1996; Jordan et al. 2009). Early problems in low-income children's mathematical understanding in the elementary and middle school grades significantly influence later school outcomes (Duncan et al., 2007), including children's ability to complete advanced mathematics coursework in high school, a gatekeeper needed to access STEM careers (National Council of Teachers of Mathematics, 2000). Given the significant wage premium of STEM careers (Deming & Noray, 2020), unequal access for children from economically disadvantaged backgrounds can effectively inhibit socioeconomic mobility and reinforce social inequality (Carter, 2006).

To address these inequities, many high-poverty school districts offer summer school as an important policy tool for remediation and grade level retention for low-achieving students (examples include Chicago, New York, and Boston; see, for example, Jacob & Lefgren, 2004; Mariano & Martorell, 2013; Matsudaira, 2008). These programs mostly focus on reading and mathematics, supported in part by research indicating that extending school time can be an effective way to support student learning for those most at risk of school failure (Patall, Cooper, & Allen, 2010).

Despite the ubiquity of summer school and the national imperative to close the opportunity gap in mathematics (National Council of Teachers of Mathematics [NCTM], 2012), we lack contemporary evidence synthesizing the characteristics of effective summer programs in mathematics. In the current article, we use data from 37 experimental and quasi-experimental studies to conduct a meta-analysis of the impacts of summer mathematics program, with the goal of understanding what program content, activities, and formats were linked to stronger student outcomes.

This review is needed for several reasons. First, although there is a long tradition of scholarly interest in summer school and the effects of seasonal school closures on inequality (e.g., Alexander et al., 2007, Atteberry & McEachin, 2020; Heyns, 1978, Kuhfeld, 2019; Quinn & Le, 2018; von Hippel, Workman, & Downey, 2018), until relatively recently, scholars rarely produced rigorous evaluations of academic summer programs, instead often using weak methodological designs that were then reflected in early syntheses of the literature (Cooper, 2000). In recent decades, scholars have produced a rich set of new studies that use stronger research designs that better support causal inference, including large-scale randomized trials (e.g., McCombs et al., 2014; Chaplin & Capizzano, 2006; Snipes et al., 2015) and regression discontinuity designs (e.g., Jacob & Lefgren, 2004; Mariano & Martorell, 2013). These interventions also include components that did not exist in summer learning programs from previous decades, such as the provision of online and blended learning and information on so-called ‘non-cognitive’ or social-behavioral outcomes. In addition, these newer studies often provide more information about program implementation than was presented in earlier reports, allowing us to examine moderators of program impact in greater detail.

This work is particularly timely given the impacts of COVID-19. While estimates of educational impacts of the COVID-19 pandemic to date have varied (e.g., Kuhfeld et al., 2020; Pier et al., 2021), it is generally acknowledged that inequity has been exacerbated and that substantial efforts are needed to help low-income students recover (Darling-Hammond, Schachner, & Edgerton, 2020). Summer school is a key policy mechanism for addressing these learning disruptions, with the American Rescue Plan Act of 2021 having allocated $29 billion for "planning and implementing activities related to summer learning and supplemental afterschool programs, including providing classroom instruction or online learning during the summer months."

Theoretical Background and Policy Context

During summer vacation, the school resources 'faucet' is turned off, and families must occupy 2-3 months of their children's time using their own means (Borman, Benson, & Overman, 2005). As a result, children's summer time use is patterned by family resources, with low-income children having fewer learning and enrichment opportunities. The U.S. average reported weekly summer program tuition is $288 -- close to 40% of household income for a free lunch-eligible family of four (Afterschool Alliance, 2015). Children from more socioeconomically advantaged families are more likely to participate in summer camps and enrichment activities, whereas low-income children are disproportionately exposed to TV (Burkam, Ready, Lee, & LoGerfo, 2004; Gershenson, 2013). Sociological research documenting the summer experiences of children from divergent socioeconomic backgrounds has found that high- and low-SES children have markedly different learning opportunities over the summer, with high-SES children generally enjoying more and better-quality summer camps, lessons, and programs that appeared to broaden both their academic knowledge and cultural capital (Chin & Phillips, 2004).

A long tradition of research in the social sciences has been concerned with the potential for seasonal school closures to exacerbate inequalities in children's learning opportunities. Early research studies comparing children's learning trajectories across seasons often indicated that low-income children were disproportionately affected (e.g., Downey, von Hippel, & Broh, 2004; Heyns, 1978), and that summer learning disparities may contribute to long-run gaps in achievement and attainment (e.g., Alexander, Entwistle, & Olsen, 2007). These early studies raised important questions for the field, yet often examined limited geographic regions or grade levels and/or used test scores that were not vertically linked (von Hippel & Hamrock, 2019). More recent research has posited the sensitivity of conclusions about summer gap widening to, for example, choice of parallel growth versus lag score models (Dumont & Ready, 2020; see also e.g., Quinn, 2015). The issue of measuring summer learning and parsing its potential contribution to inequality remains an active source of scholarly debate (see e.g., von Hippel, 2019; and Alexander, 2019). However, there is general agreement among scholars that children learn reading and math more slowly during the summer than during the school year, and that summer therefore affords children opportunities, whether to catch up if they are struggling or to gain new knowledge if they already progressing well (e.g., von Hippel, 2019).

In response to these issues, many school districts have adopted summer learning programs to advance remediation and equity goals. Until recently, relatively few of these programs were rigorously evaluated. The most recent review of the empirical literature is over two decades old (Cooper et al., 2000), and includes studies lacking in methodological rigor, such as pre-post study designs that do not account for children's normal maturation. Although Cooper et al. (2000) recognize this limitation and try to adjust for variation in study quality, they note significant differences in the magnitude of effects depending on certain methodological features. Lauer et al. (2006) report a similar finding in their meta-analysis of the effects of out-of-school-time programs on student achievement, suggesting the importance of using more rigorous methods to more accurately understand the effectiveness of summer programs.

In recent decades, growing dedication among funders and researchers to research designs that support causal inference has yielded a larger set of higher quality studies on summer math programs, including large-scale RCTs and RDs conducted in multiple large U.S. urban districts. The results of these studies appear somewhat variable. For example, a large-scale randomized experimental evaluation of the BELL summer program reported mostly non-significant findings on reading, math, and social-emotional outcomes (Somers, Welbeck, Grossman, & Gooden, 2015). RAND conducted a large-scale, randomized experimental study of voluntary summer programs in five U.S. districts, and found significant and positive impacts on math scores, but not on reading scores or social-emotional skills. In the current review, we aim to identify features of these interventions that explain this variability.

Lastly, to our knowledge, no prior synthesis has examined the impacts of summer school programs on children's social-emotional and behavioral outcomes. Despite the relatively small number of studies reporting these outcomes, compiling the emerging evidence is important given well-documented income gaps in these outcomes (e.g., Downey, Workman, & von Hippel, 2019), and concerns that academic-focused summer programs may have unintended negative effects on social-emotional development by, for example, taking students away from leisure pursuits during break.

In summary, the current review synthesizes the recent empirical literature on summer mathematics learning programs in order to understand what program characteristics and contextual factors are associated with stronger student academic and social-emotional and behavioral outcomes.

Method

We investigate these questions by conducting a comprehensive meta-analysis of the experimental and quasi-experimental literature on the impacts of summer learning programs in mathematics. Meta-analysis is well-suited to this aim because it allows us to pool information across multiple studies, and to examine multiple hypothesized moderators of program impact.

Search Procedures

This review captures information on summer programs that aim to boost children's academic achievement in mathematics, including both mandatory programs, such as district-required summer school programs for children who have failed the previous grade, as well as optional programs, which parents may choose for children's enrichment. Summer programs are either classroom-based, with children attending in person, or home-based, with math activities given to the child to complete alone or with family members.

We developed a database of studies via a four-phase search process. We searched these channels from August 1998, as this was the last date for which searches were conducted in the prior meta-analysis of the summer school literature (Cooper et al., 2000). Searches were completed through April 2020. In the first search phase, we conducted an electronic search using the databases Academic Search Premier, Education Abstracts, ERIC, PsycINFO, EconLit, and ProQuest Dissertations and Theses, for the period August 1998 through April 2020. Searches were conducted using subject-related keywords relating to summer programs and methodology-related keywords designed to capture experimental and quasi-experimental designs adapted from Kim & Quinn (2012). Second, we searched targeted internet sites including the What Works Clearinghouse, Harvard Family Research Project's Out-of-School Time Database, MDRC, NBER, RAND, AIR, Mathematica, Wallace Foundation, and the National Summer Learning

Association. We also searched the abstracts of the Society for Research on Educational Effectiveness (SREE) conference. Third, we scanned the reference lists of previous review articles (Alexander, Pitcock, & Boulay, 2016; Bodilly & Beckett, 2005; Lauer et al., 2006; McCombs et al. 2011; McCombs et al., 2019; Terzian, Moore, & Hamilton, 2009). Lastly, we contacted government agencies requesting relevant research reports.

The search procedures described above yielded 2,544 records identified via database screening, and an additional 15 records identified through other sources (see Figure 1 for screening flowchart). After removing duplicates, we were left with 1,958 records.

Study Inclusion Criteria

The studies included in our review needed to be published after August 1998 and meet the following five selection criteria:

(1) Evaluate the effects of a classroom- or home-based summer mathematics intervention;

(2) Present mathematics learning outcomes for treatment and control groups of students;

(3) Include students who were entering pre-K-12 prior to enrollment in a summer math intervention;

(4) Compare the performance of students in a treatment group to the performance of students in a control group who did not participate in the treatment or systematically receive an alternative intervention; and

(5) Provide sufficient empirical information to compute an effect size (Hedges' *g*-index).

We also required that studies provide evidence that the achievement levels of treatment and control groups were comparable at baseline, as discussed below. We admitted studies that used randomized experimental and regression discontinuity designs, as well as studies in which scores were matched at pretest through exact matching or propensity score matching. If pretest differences were between 0.05 and 0.25 SD, we required that the authors had performed statistical adjustments for pretest differences (e.g., ANCOVAs); in cases where these were not presented in study reports, we manually calculated a difference in differences adjustment by subtracting the standardized pre-intervention difference from the standardized difference in outcomes, per What Works Clearinghouse guidelines (WWC, 2020).

Study Screening

Screening was conducted in two phases. First, two raters screened each of the studies' titles and abstracts to identify potentially relevant studies, flagging studies that met or potentially met criteria 1-4. All studies flagged as potentially relevant by either rater were reviewed by one of the authors, who made a final determination about moving the study forward. A total of 104 studies met these initial relevance criteria and advanced to full-text screening.

In the second screening phase, two raters working independently, including at least one study author, examined the full text of each study and applied a more detailed set of methodological inclusion criteria. We required that studies present sufficient information to calculate an effect size, along with evidence that the treatment and control groups' achievement levels were comparable at baseline. We judged studies to have met this standard that used randomized experimental and regression discontinuity designs, as well as studies in which scores were matched at pretest through exact matching or propensity score matching. If pretest differences were between 0.05 and 0.25 SD, we required that the authors had performed statistical adjustments for pretest differences or manually adjusted the effect sizes as discussed above. We excluded summer programs with no mathematics learning component, such as summer social skills or book reading programs. We required that participating students were entering grades pre-K-12; because of our conceptual interest in summer learning during seasonal school closures, we excluded studies that examined only preschool children who had no formal schooling prior to their summer program participation. The most common reasons for exclusion were for characteristics of the intervention (e.g., off-topic, did not evaluate the effects of a classroom- or home-based summer mathematics intervention; N=17), methodological issues (e.g., no control group, N=6; no pretest data or pretest data not equivalent at baseline; N=19), and lack of outcome data (i.e., did not present student outcomes for treatment and control; N=14). Note that some studies had multiple reasons for exclusion (see Figure 1).

Thirty-seven studies met the full review inclusion criteria and advanced to study coding. In situations where study authors produced multiple reports on the same study, we used all available study documents to glean information about the intervention and study impacts, and used the most recent version (often the peer-reviewed version) as authoritative in case of any inconsistencies across reports. Of the included studies, 43% were published or unpublished doctoral dissertations; 38% were technical reports including contract researchers' reports and district, state, or federal government reports; and 19% were peer-reviewed journal articles. Many of these studies contributed multiple effect sizes due to multiple outcome measures, multiple samples, multiple versions of the same program with a common control group, and/or multiple programs.

The final meta-analytic sample includes 164 effect sizes nested within these 37 studies. The sample includes a separate effect size for each treatment contrast, each assessment of mathematics achievement, other academic attainment, and social-emotional and behavioral outcomes, and each sample of students that the study reported. As discussed below, we used the robust variance estimation (RVE) approach (Tanner-Smith & Tipton, 2014) to account for the nested nature of the data.

Study Coding

Study authors and trained graduate research assistants conducted full-text coding using the following procedures. Before beginning double-coding, we established inter-rater reliability. Each week, each member of the team coded two studies, then met to reconcile disagreements and refine codebook descriptions. We repeated this procedure until we reached a stable set of codes and an 80% agreement threshold. Each study was then coded by two researchers, including at least one study author. Each researcher coded the studies independently, then raters met to reconcile discrepant codes. All disagreements were resolved through discussion.

Effect Sizes Calculation

Standardized mean difference effect sizes were calculated using Hedges' g:

$$g = J \times \frac{(\overline{Y_E} - \overline{Y_C})}{S^*}.$$

In this formula, $\overline{Y_E}$ represents the average treatment group outcome, Y_C represents the average control group outcome, and S^* represents the pooled within-group standard deviation. J represents a correction factor that adjusts the standardized mean difference to avoid bias in small samples:

$$J = 1 - \frac{3}{4 \times (N_E + \mathrm{N_c} - 2) - 1},$$

In this equation, N_E represents the number of students in the treatment group and N_C represents the number of students in the control group. Effect sizes were calculated using the software package Comprehensive Meta-Analysis (CMA) in the majority of cases.

Model Selection

A frequent problem in meta-analysis occurs when a single study presents multiple effect sizes, because study authors often measure interventions' impacts on several different outcomes.

These effect sizes nested within a single study are likely to be correlated, which violates the assumption of statistical independence. To address this issue, prior meta-analyses of the impacts of summer programs have either averaged effect sizes, or selected a single effect size per study to ‘represent’ that study in the analysis (e.g., Kim & Quinn, 2014). However, we argue that a robust variance estimation (RVE) approach detailed in Tanner-Smith and Tipton (2014) more appropriately models our data. This approach adjusts standard errors to account for the dependencies among effect sizes within the individual studies, akin to adjusting standard errors in ordinary least squares (OLS) regression models for heteroscedasticity (e.g., using Huber–White standard errors) or to account for the nesting of data within clusters (e.g., clustered standard errors). Importantly, this approach permits us to include multiple effect sizes from a single study in our analysis. In this way, it overcomes the losses of information common in traditional meta-analyses that could only accommodate a single effect size per study, and thus either dropped effect sizes or included a single average effect size for each study (see Tanner-Smith & Tipton, 2014). RVE models can also control for methodological features known to influence outcomes (e.g., C. J. Hill et al., 2008; Taylor et al., 2018). This approach has been used in several recent meta-analyses where individual studies report multiple effect sizes (e.g., Clark, Tanner-Smith, & Killingsworth, 2016; Dietrichson, Bøg, Filges, & Klint Jørgensen, 2017; Gardella, Fisher, & Teurbe-Tolon, 2017; Lynch et al., 2019).

The RVE modeling approach was developed to account for two kinds of dependencies among effect sizes. The first type of dependency is *correlated effects*, which occurs when a study presents multiple effect size estimates for a single underlying construct or of correlated underlying measures, or uses the same control group for multiple treatment contrasts. The second type is *hierarchical effects*, which arises when multiple treatment-control contrasts are nested within a larger cluster of experiments (e.g., a single research team conducts several evaluations of the same program). When both types of dependencies are present, Tanner-Smith and Tipton (2014) recommend selecting a method based on the more frequent type of dependence. As correlated effects predominated in our data, we used the recommended inverse variance weights recommended by Tanner-Smith and Tipton (2014).

The weight for effect size *i* in study *j* is calculated using the following formula:

$$w_{ij} = \frac{1}{\{(v_{*j}+\tau^2)[1+(k_j-1)\rho]\}},$$

where $v_{*j}$is the mean of within-study sampling variances $\left(SE_{ij}^2\right)$ within each study, τ^2 is the estimate of the between-studies variance component, k_j is the number of effect sizes within each study, and ρ is the assumed correlation between all pairs of effect sizes within each study. Thus effect sizes from studies contributing more effect sizes and with higher sampling variances are given lower weight. It is assumed that ρ is constant across studies, and we use the recommended default value of ρ = .80 (Tanner-Smith & Tipton, 2014). However, simulation studies suggest that results using the RVE approach are not sensitive to values of ρ (e.g., Tanner-Smith & Tipton, 2014; S. J. Wilson, Tanner-Smith, Lipsey, Steinka-Fry, & Morrison, 2011). We also conduct a series of sensitivity analyses to gauge whether our results are sensitive to alternative values of ρ. We use the *robumeta* package in Stata 15 (developed by Tanner-Smith & Tipton, 2014) to estimate our RVE models and incorporate the small-sample correction proposed by the developers of the RVE approach (Tanner-Smith & Tipton, 2014; Tipton & Pustejovsky, 2015). We also report the results of *F* tests to test the joint significance of the program features included in our RVE models. These *F* tests were conducted using the *robumeta* and *clubSandwich* packages in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

As described above, we grouped potential moderators of program impact into four categories indexing the time/duration, focus, activities, and resources of summer programs. To examine whether specific features in each category moderated program impact, we fit four sets of conditional meta-regression models with RVE, including the coded features as moderators and treating these moderators as fixed. Within each category, we first modeled the effect of each code separately. To further understand their joint relationships, we then fit a model entering all codes together. All models controlled for whether the study was an RCT or RD as a measure of design quality; further indicators of design quality are discussed below.

Moderator Analyses

We coded all studies for four categories of potential moderators: (1) *study design and sample characteristics*; (2) *program foci*; (3) *program activities*; and (4) *program resources*. We coded each study on a set of methodological criteria, noting whether the study design was a randomized experiment or regression discontinuity design versus another type of quasi-experiment. We noted the type of publication (e.g. peer-reviewed journals, technical reports, dissertations). In addition, we classified each outcome into one of three types. Most effect sizes captured students' mathematics achievement on *standardized tests*, including state tests and large national assessments (e.g., NWEA, ITBS). A number of studies also captured information on *other mathematics attainment outcomes*, including mathematics course grades, math course-taking, and completing a STEM degree. Lastly, several studies reported summer program impacts on students' 'non-cognitive' or *social-emotional and behavioral outcomes*, including outcomes such as social skills, attendance, absenteeism, disciplinary referrals, suspensions, and affective domains such as mathematics interest and self-efficacy.[1] We operationalized the poverty level of the samples using data from study reports on the percentage of students eligible for a free or reduced-price school lunch (or in poverty by another indicator).

We developed three categories of codes to capture summer program characteristics. The first set of codes examined the summer program's *focus*. We classified each program as focused on either mathematics only, or as possessing a broad academic focus, including other academic subjects (e.g., reading, science, social studies) in addition to mathematics. We classified the goals of each program as broadly focused on either remediation or on future coursework for the next grade level via preparation and/or preview of future content. We captured whether the summer program was conducted fully online, versus in-person. Lastly, we evaluated whether each program's content as described was aligned with the National Council of Teachers of Mathematics (NCTM) and/or Common Core State Standards (CCSS) professional teaching standards in mathematics.

A second set of codes examined the *activities* in which children participated during the summer program. We coded each study for evidence that children participated in hands-on projects, completed textbook exercises, engaged in group work, and/or completed computer-based skills practice, over the course of the summer program. We also included in our analyses a composite index of the total number of these activities that were reported per study.

A third set of codes indexed the *resources* available at each summer program. First, we captured information about the *duration* of the program, using codes for program hours per day, the total program hours, and the number of program hours per day spent specifically on

[1] We did not require demonstration of outcome-specific baseline equivalence for social-behavioral outcomes. Pretest data for these outcomes was often not reported, and for certain outcomes (e.g., dropout) there is no directly corresponding baseline variable.

mathematics. We coded each study for information about summer program *staffing*, including evidence that the summer program instructors received professional development, as well as whether teachers received explicit direction in preparing for summer instruction, such as pre-made lesson plans. We examined *district and community supports*, including whether summer programs provided transportation for students (such as bus rides), and whether the program provided free meals on site (breakfast and/or lunch). Lastly, we captured information about the average *class size* in the summer program.

Results

Descriptives

Table 1 presents descriptive statistics regarding the studies and summer programs included in our dataset. The study designs employed included a mix of randomized experiments and regression-discontinuity designs (30% of studies), including several large-scale randomized trials and RD studies conducted in large, high-poverty urban school districts (e.g., Jacob & Lefgren, 2004; Mariano & Martorell, 2013; McCombs et al., 2020), along with propensity score matching and other quasi-experimental designs that demonstrated satisfactory group equivalence at baseline, as described above (70% of studies). These reports included evaluations of large programs such as Upward Bound Math and Science (Olsen et al., 2007) and the Building Educated Leaders for Life (BELL) program (Somers et al., 2015), along with smaller programs conducted at the school- and district-level.

Table 1 also presents study-level frequencies for included studies' summer programs' characteristics, including foci, activities, and resources. Most summer programs evaluated were conducted in-person (89%), while 11% of studies evaluated fully-online programs. Median summer program duration was 90 hours. The programs examined in our dataset primarily served low-income students: On average, among studies with available data, 64% of children in the included study samples were eligible for a free or reduced price school lunch. All in-person programs in our dataset that provided information on teachers' qualifications were taught by either certified teachers or a mix of teachers and aides. Most programs (79%) were focused on remediation of previous years' academic content; most were also broad-based in academic focus, with 79% offering instruction on a range of other academic subjects in addition to mathematics. Approximately a third of programs (32%) reportedly used curriculum materials or activities aligned to Common Core and/or NCTM standards. Most studies (54%) reported that instructors received professional development prior to or during the summer, while 24% of studies reported that specific lesson plans or structures were provided. Among studies reporting class size data, average class size was 17 students.

Overall Average Impacts

Compared to control group children, children who participated in summer programs that included mathematics lessons and activities enjoyed significant improvements in mathematics learning as well as social-behavioral outcomes. We find an average weighted impact estimate of +0.09 standard deviations on mathematics achievement outcomes (see Table 2). Restricting the outcomes to standardized mathematics achievement tests, we find an average weighted impact estimate of +0.10 standard deviations. To contextualize the magnitude of this effect, a typical treatment group student who participated in a summer program would be expected to rank approximately 4 percentile points higher than a typical control group student (Lipsey et al., 2012). Summer programs also showed average positive impacts on students' social-emotional learning (SEL) and behavioral outcomes, with average weighted impact estimate of +0.09 standard deviations. Pooled across outcome types, of the 164 effect sizes included in the meta-analysis, 110 were positive in sign (67%), and 44 of these were statistically significant. Forty-eight were negative in sign, with three of these being statistically significant. Six effect sizes had point estimates of zero.

There were no significant differences in average effect sizes based on outcome measure type (i.e., standardized math achievement tests, broader school math attainment outcomes, or social-emotional/behavioral outcomes); study publication type (e.g., peer-reviewed journal versus not); or student grade level (elementary versus middle/high school), or whether the effect sizes in the study were adjusted for covariates. Study design (randomized experiment/regression discontinuity versus other designs) was also not a significant predictor of effect size magnitude

Features That Moderate Program Impacts

We next examine factors that may moderate impacts on student outcomes, beginning with sample characteristics. In the subsequent analyses, we restrict the outcomes of interest specifically to mathematics achievement outcomes; we exclude social-behavioral outcomes from subsequent moderator analyses both because they were presented in comparatively few studies (7) and because conceptually, the factors that would be expected did not find a significant relationship between the poverty level of the student sample and program impacts (see Table 3). For this analysis, we operationalized poverty level using a continuous indicator for the proportion of children in the sample who were eligible for free or reduced-price school lunch. We also did not find significant differences when operationalizing poverty level as a dichotomous indicator (i.e., high-poverty defined as including 75 percent or greater of children eligible for a free or reduced-price school lunch; 49% of included samples). These results suggest that summer programs have average positive impacts for children in both high-poverty and relatively lower-poverty contexts.

We turn next to summer program duration. We did not find significant relationships between total program hours or program hours per day and students' math outcomes. However, we did find that average effect sizes were larger when programs spent more hours per day on mathematics (+0.09 SD, $p < .10$) (see Table 4).

Next, we examined the associations between the focus of the summer program and effect sizes via a series of multilevel regression models (Table 5). Average effect sizes were larger when programs focused specifically on mathematics, as compared with having a broader focus on multiple academic subjects (+0.17 SD, $p < 0.05$), and this result remained significant in the final model. A focus on remediation, as compared with preparation for future coursework, was negatively associated with effect size magnitude in the final model that included all predictors (-0.11 SD, $p < 0.10$). The inclusion of content judged to be aligned with NCTM and/or CCSS standards in mathematics was not a significant predictor of effect size magnitude. Descriptively, programs that were fully online had smaller impacts on average than did fully in-person programs, although as noted above, the number of fully-online programs was relatively small, and this relationship was not statistically significant.

We then turned to the relationships between effect sizes and summer program activities (Table 6). We found that the use of textbook exercises was negatively associated with effect size magnitude, and this relationship retained its significance in the final model (-0.12 SD, $p < 0.05$). No other program activities for which we coded – use of a commercially-available curriculum, hands-on projects, group work, or computer-based skills practice -- were significantly associated with effect size magnitude.

Table 7 examines the relationships between summer program *resources* and effect sizes. Of the activities for which we coded – the provision of teacher professional development, teacher direction in lesson planning, including the provision of pre-written lesson plans, transportation for students, and average class size – none were significantly associated with effect size magnitude, either individually or in the combined model.

Discussion and Conclusion

In summary, we found that studies of summer programs in mathematics had, on average, positive effects on both student achievement and social-emotional and behavioral outcomes, with an average pooled effect size across studies of +0.09 standard deviations. Summer programs had average positive impacts on both standardized math tests (+0.10 SD) and social-emotional and behavioral outcomes (+0.09 SD).

To contextualize the magnitude of these achievement impacts, researchers have estimated that a typical teacher who raises student achievement on standardized tests by +0.14 SD creates marginal gains of roughly US$7,000 per child in present value future earnings (Chetty, Friedman, & Rockoff, 2014). Extrapolating from this, the estimated average test score impact of summer programs of +0.10 SD would be expected to yield approximately US$5,000 in present value future earnings per student. The weighted mean impact estimate of summer programs constitutes a 'medium' effect, as classified using recent benchmarks for studies of pre-K–12 education interventions evaluating effects on student achievement (Kraft, 2020). Another relevant benchmark is the cost-benefit ratio of summer school as an investment. Relatively few studies provided detailed budget data. However, a useful example adapted from Matsudaira (2008) suggests a possible comparison. Examining the results of the Tennessee STAR experiment, Krueger found that reducing class size by one-third improved student achievement by 0.22 standard deviations, at an estimated cost (in current dollars) of $6,600 per student. By contrast, the per student budget of the summer school in a large, high-poverty urban district represented in our study sample (in current dollars) was $1900 (Matsudaira, 2008). These estimates suggest that the cost-benefit payoff of summer school may be more than 40% greater than a class-size reduction in terms of boosting student achievement.

We conducted a series of analyses to examine the relationships between summer program characteristics and the magnitude of student outcomes. The characteristics that were significantly associated with effect size magnitude included the following:

- More program hours per day spent on mathematics;
- A mathematics-specific program focus;
- A focus on previewing future mathematics content (as opposed to remediation).

The finding that programs specifically focused on mathematics garnered large impacts on math achievement may be unsurprising, given these programs' targeted nature. Programs that targeted a broader variety of academic subjects, such as reading and social studies, may well have produced academic benefits in other subject areas that we did not capture here, given our specific interest in mathematics. Although it is possible that programs focused on previewing future mathematics content produced larger impacts because they enrolled learners who were interested in such enrichment, and had less need of remediation, another possibility is that previewing the next year's content is simply more helpful to students than revisiting prior material. In the United States, mathematics textbooks and instruction are already quite redundant, with a substantial amount of content retaught across multiple grades and a heavy existing focus on reviewing old material (Polikoff, 2012). This tendency persists despite the fact that policy documents have argued that much of this review time should be reallocated to developing new material in greater depth (Romberg, 1989). Perhaps summer programs that focused on presenting new material were better able to capture students' interest, avoid the boredom of repeating old content, and deepen students' understanding of the next grade's material.

Lastly, we know of no prior reviews that have synthesized the impacts of summer programs on social-emotional and behavioral outcomes. Summer is often considered a time of leisure, and policymakers sometimes express the concern that summer school will worsen children's affective dispositions or behaviors in the subsequent school year. We find no evidence to substantiate this concern. On the contrary, in our meta-analysis we found that summer programs improved students' social-emotional and behavioral outcomes. This finding points to the promise of summer programs to remediate learning gaps while not harming, but bolstering, students' 'non-cognitive' skills, social-emotional learning, and behaviors.

Preliminary Conclusions

Summer programs are a malleable factor to reduce inequities in children's mathematics learning. Our analysis found that summer programs had positive impacts for children in both higher- and lower-poverty contexts, and in both achievement and non-achievement domains. In the current policy context, summer school may be a cost-effective mechanism for school districts in remediating learning problems resulting from the COVID-19 disaster. Contemporary summer programs can help to bolster children's mathematics learning, paving the way for strengthened long-run STEM educational opportunities and outcomes.

References

Alexander, K. L. (2019). Summer learning loss sure is real. *Education Next*. Retrieved from https://www.educationnext.org/summer-learning-loss-sure-is-real-response/

Alexander, K. L., Entwisle, D.R., Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, 23(2), 171–191.

Borman, G. D., Benson, J., Overman, L. T. (2005). Families, schools, and summer learning. *The Elementary School Journal*, 106(2), 131–150.

Chaplin, D., & Capizzano, J. (2006). Impacts of a Summer Learning Program: A Random Assignment Study of Building Educated Leaders for Life (BELL).

Chin, T., & Phillips, M. (2004). Social reproduction and child-rearing practices: Social class, children's agency, and the summer activity gap. *Sociology of Education*, 77(3), 185–210.

Cooper, H., Charlton, K., Valentine, J. C., Muhlenbruck, L., & Borman, G. D. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monographs of the Society for Research in Child Development*, i-127.

Dorn, E. Hancock, B., Sarakatsannis, J. & Viruleg, E. (2020). *COVID-19 and student learning in the United States: The hurt could last a lifetime*. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-and-student-learning-in-the-united-states-the-hurt-could-last-a-lifetime

Dumont, H., & Ready, D. D. (2020). Do schools reduce or exacerbate inequality? How the associations between student achievement and achievement growth influence our understanding of the role of schooling. *American Educational Research Journal*, *57*(2), 728-774.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Entwisle, D. R., Alexander, K. L., Olson, L. S. (2001). Keep the faucet flowing: Summer learning and home environment. *American Educator*, 25(3), 10–15, 47.

Gershenson, S. (2013). Do Summer Time-Use Gaps Vary by Socioeconomic Status? *American Educational Research Journal*, *50*(6), 1219–1248. https://doi.org/10.3102/0002831213502516

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of economics and statistics*, *86*(1), 226-244.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental psychology*, *45*(3), 850–867. https://doi.org/10.1037/a0014939

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Krueger, A., 2002. Economic considerations and class size. National Bureau of Economics Research Working Paper.

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E., & Liu, J. (2020). Projecting the Potential Impact of COVID-19 School Closures on Academic Achievement. Educational Researcher, 49(8), 549–565. https://doi.org/10.3102/0013189X20965918

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

Mariano, L. T., & Martorell, P. (2013). The academic effects of summer instruction and retention in New York City. *Educational Evaluation and Policy Analysis*, *35*(1), 96-117.

McCombs, J. S., Pane, J. F., Augustine, C. H., Schwartz, H. L., Martorell, P., & Zakaras, L. (2014). *Ready for fall? near-term effects of voluntary summer learning programs on low-income students' learning opportunities and outcomes*. Rand Corporation.

McCombs, J. S., Augustine, C. H., Unlu, F., Ziol-Guest, K. M., Naftel, S., Gomez, C. J., ... & Todd, I. (2019). *Investing in Successful Summer Programs: A Review of Evidence under the Every Student Succeeds Act. RR-2836-WF*. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138.

National Council of Teachers of Mathematics [NCTM]. (2012). Closing the Opportunity Gap in Mathematics Education: A Position of the National Council of Teachers of Mathematics.

Retrieved from https://www.nctm.org/uploadedFiles/Standards_and_Positions/Position_Statements/Opportunity%20Gap.pdf

Patall, E. A., Cooper, H., & Allen, A. B. (2010). Extending the school day or school year: A systematic review of research (1985–2009). *Review of educational research*, *80*(3), 401-436.

Pier, L., Hough, H. J., Christian, M., Bookman, N., Wilkenfeld, B., Miller, R., (2021). *COVID-19 and the Educational Equity Crisis: Evidence on Learning Loss from the CORE Data Collaborative*. https://edpolicyinca.org/newsroom/covid-19-and-educational-equity-crisis

Quinn, D. M. (2015). Black–White summer learning gaps: Interpreting the variability of estimates across representations. *Educational Evaluation and Policy Analysis*, *37*(1), 50-69.

Quinn, D. M., & Le, Q. T. (2018). Are we trending to more or less between-group achievement inequality over the school year and summer? Comparing across ECLS-K cohorts. *AERA Open*, *4*(4), 2332858418819995.

Snipes, J., Huang, C. W., Jaquet, K., & Finkelstein, N. (2015). The Effects of the Elevate Math Summer Program on Math Achievement and Algebra Readiness. REL 2015-096. *Regional Educational Laboratory West*.

U.S. Department of Education, National Center for Education Statistics. (2020). *The Condition of Education 2020* (NCES 2020-144).

von Hippel, P. T. (2019). Is summer learning loss real? How I lost faith in one of education research's classic results. *Education Next*, *19*(4), 8-15.

von Hippel, P. T. (2019). Summer Learning: Key Findings Fail to Replicate, but Programs Still Have Promise. *Education Next.*

Tables and Figures

Table 1

Categories and descriptions of codes

Code	Code description	Code present[a]
Intervention Type		
In-person	Study examined a summer program conducted in-person at a school or other out-of-home location.	89%
Online only	Study examined a summer program conducted exclusively online.	11%
Research Design and Sample Characteristics		
RCT or RD	Study used a randomized controlled trial or regression discontinuity design.	30%
Publication type	Study is a dissertation.	43%
	Study is a peer-reviewed journal publication.	19%
	Study is a technical reports including contract researchers' reports and district, state, or federal government reports.	38%
Grade level – Elementary	Study sample included elementary (pre-K-5) students (versus middle/high school).	46%
Poverty level	Percentage of students reported eligible for free or reduced-price school lunch.	64%
Effect Size Type		
Standardized mathematics test outcome	Outcome is a standardized mathematics achievement test (either state or nationally-standardized [e.g., NWEA MAP, ITBS)].	86%
School attainment outcome	Outcome is an academic outcome besides test score (e.g., course grades).	14%
Social-emotional/behavioral outcomes	Outcomes is a social-emotional or behavioral outcome.	23%
Adjusted for covariates	Effect size is adjusted for covariates (e.g., pretest score).	92%

Code	Code description	Code present[a]
Summer Program Focus		
Math-specific focus	The summer program focuses specifically on mathematics, in contrast to broad-based programs that also included other academic subjects (e.g., reading, science, social studies).	22%
Program Goals	The summer program focused on remediation, learning loss, or 'catch up.'	78%
	The summer program focused on future coursework or the next grade level via preparation and/or preview of future content.	22%
Standards alignment	The summer program content was aligned with NCTM and/or CCSS standards.	32%
Summer Program Activities		
Children's activities	Variables indexing children's participation in hands-on projects, textbook exercises, group work, and computer-based skills practice, as well as total number of activities reported (range 0 – 3).	0.92
Curriculum: Commercial program	The study reported the summer program's use of a commercially-available curriculum.	27%
Summer Program Resources		
Staffing	Program instructors received PD, either prior to or during the summer.	54%
	Teacher direction (lesson plans or structure) was provided.	27%
District/community supports	Transportation was provided.[b]	47%
	Meals (breakfast and/or lunch) were provided.	46%
Average class size	Average class size	16.6

Note: N = 37 studies.
[a] Figures in the third column include the percent of studies which feature the row code for binary variables, or the sample average calculated at the study level for continuous variables (e.g., contact hours). For studies that had the feature present in one treatment arm but not another treatment arm, the code is counted as present if it is present in any treatment arm.
[b] Conditional on the summer program being offered in-person.

Table 2

Results of estimating unconditional meta-regression models with robust variance estimation (RVE)

	Dependent variable: All Math Outcomes Effect Size (Hedges's g)	Dependent variable: Standardized Math Achievement Tests Effect Size (Hedges's g)	Dependent variable: Social-Behavioral Skills Effect Size (Hedges's g)
Constant	0.086***	0.095***	0.086*
	(0.022)	(0.024)	(0.034)
N effect sizes	164	112	38
N studies	37	37	7
τ^2 [a]	0.008	0.008	0.007
95% prediction interval[b]	(-0.084, 0.256)	(-0.077, 0.267)	(-0.075, 0.247)

Note: We assume the average correlation between all pairs of effect sizes within studies is 0.80.
[a] τ^2 is the method of moments estimate of the between study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014). [b] The 95% prediction interval is calculated as the estimated average effect size +/- 1.96* τ.
*p < .10, **p < .05, ***p < .01.

Table 3

RVE results including sample characteristics (poverty level of sample) as moderators

	Dependent variable: Effect Size (Hedges's g)
Between-study effects	
% of sample eligible for free- or reduced-price school lunch (standardized)	-0.032 (0.036)
N effect sizes	105
N studies	32
τ^{2} [a]	0.006
	Weighted mean ES: Effect Size (Hedges's g)
High-poverty sample (% low income >0.75)	0.0838***
Mid-low poverty sample (% low income<=0.75)	0.119**

Note: We assume the average correlation between all pairs of effect sizes within studies is 0.80. Models include a control for RCT or RD study design at the between-study level.
[a] τ^2 is the method of moments estimate of the between study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).
**p* < .10, ***p* < .05, ****p* < .01.

Table 4

RVE results including program duration/intensity as moderators

	Dependent variable: Effect Size (Hedges's g)		
Between-study effects			
Total program hours	0.000 (0.000)		
Program hours per day		-0.002 (0.015)	
Hours per day on math			0.090* (0.038)
N effect sizes	100	73	54
N studies	31	28	22
τ^{2a}	0.006	0.007	0.003

Note: We assume the average correlation between all pairs of effect sizes within studies is 0.80. Models include a control for RCT or RD study design at the between-study level.
[a] τ^2 is the method of moments estimate of the between study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).
**p* < .10, ***p* < .05, ****p* < .01.

Table 5

RVE results including summer program foci as moderators

	Dependent variable: Effect Size (Hedges's g)				
Between-study effects					
Math-specific Focus	0.173* (0.083)				0.160** (0.064)
Program Goal: Catch up		-0.132 (0.081)			-0.112* (0.057)
Standards-aligned			0.088 (0.058)		0.045 (0.041)
Online-only Program				-0.056 (0.105)	-0.085 (0.055)
N effect sizes	112	112	112	112	112
N studies	37	37	37	37	37
τ^2 [a]	0.007	0.008	0.009	0.009	0.007
Results of joint F-test					*F* = 2.78, *df* = 4, *p* = 0.132

Note: We assume the average correlation between all pairs of effect sizes within studies is 0.80. Models include a control for RCT or RD study design at the between-study level.

[a] τ^2 is the method of moments estimate of the between study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

**p* < .10, ***p* < .05, ****p* < .01.

Table 6

RVE results including summer program activities as moderators

	Dependent variable: Effect Size (Hedges's g)						
Between-study effects							
Commercially available curriculum	-0.017 (0.058)						
Hands-on projects		0.078 (0.073)				0.096 (0.084)	
Textbook exercises			-0.106** (0.042)			-0.112** (0.041)	
Group work				-0.005 (0.069)		-0.032 (0.078)	
Computer-based skills practice					-0.003 (0.072)	-0.003 (0.067)	
Number of summer program activities							-0.004 (0.027)
N effect sizes	112	112	112	112	112	112	112
N studies	37	37	37	37	37	37	37
τ^2 [a]	0.010	0.008	0.013	0.009	0.009	0.014	0.011
Results of joint F-test						*F* = 0.334, *df* = 4, *p* = 0.847	

Note: We assume the average correlation between all pairs of effect sizes within studies is 0.80. Models include a control for RCT or RD study design at the between-study level.

[a] τ^2 is the method of moments estimate of the between study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

**p* < .10, ***p* < .05, ****p* < .01.

Table 7

RVE results including summer program resources as moderators

			Dependent variable: Effect Size (Hedges's g)		
Between-study effects					
Teacher PD	0.007 (0.051)			-0.035 (0.066)	
Lesson Plans		0.081 (0.067)		0.090 (0.073)	
Transportation			-0.045 (0.056)	-0.043 (0.065)	
Average class size					0.014 (0.010)
N effect sizes	112	112	105	105	51
N studies	37	37	34	34	17
τ^{2a}	0.009	0.013	0.013	0.018	0.006
Results of joint F-test				F = 0.658, df = 3, p = 0.591	

Note: We assume the average correlation between all pairs of effect sizes within studies is 0.80. Models include a control for RCT or RD study design at the between-study level. Studies of online-only programs are excluded from the analysis of transportation. Average class size information was available for a subset of studies.

[a] τ^2 is the method of moments estimate of the between study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

**p* < .10, ***p* < .05, ****p* < .01.

Figure 1
PRISMA Flow Diagram