2021

Examining Bridges in Mathematics and Differential Effects Among English Language Learners

Garret J Hall, Patti Schaefer, Teri Hedges and Eric Grodsky

The Version of Record of this manuscript has been published and is available at *School Psychology Review* (2022)

https://www.tandfonline.com/doi/full/10.1080/2372966X.2020.1871304

Examining *Bridges in Mathematics* and Differential Effects Among English Language Learners

Garret J. Hall[a], Patti Schaefer[b], Teri Hedges[b], and Eric Grodsky[c]

[a]Florida State University

[b]Madison Metropolitan School District

[c]University of Wisconsin—Madison

Author Note

We have no known conflicts of interest to disclose.

Correspondence regarding this manuscript should be sent to Garret J. Hall, Assistant Professor, Department of Educational Psychology and Learning Systems, Florida State University, 1114 W. Call St, Office 3204H, Tallahassee, FL 32306, email: gjhall@fsu.edu

We thank Dr. Silvia Romero-Johnson and Ben Kollasch in Madison Metropolitan School District for additional help on this project. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Award #R305B150003 to the University of Wisconsin—Madison and made possible by the Madison Education Partnership. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Abstract

Determining the effectiveness of core mathematics curricula is foundational to evidence-based practice. Examining effectiveness heterogeneity is also crucial to better understanding mathematics achievement among English language learners (ELLs). In this study, we used a quasi-experimental design (difference-in-differences) to examine the impact of a standards-based elementary mathematics curriculum (*Bridges in Mathematics*) on fifth graders' annual gains in mathematics achievement in a large midwestern school district compared to the district's prior curriculum (*Investigations*). We also investigated whether the effect of *Bridges* varied across English language proficiency (ELP) levels of English language learners (ELLs). Students in schools that implemented *Bridges* (n = 1,839) showed significantly greater mathematics gains compared to those receiving the prior curriculum (n = 3,354; g = 0.25 in change score standard deviations). This effect did not vary significantly across ELP levels. Limitations of this study as well as implications for research and practice with core curricula are discussed.

Keywords: Elementary school, mathematics, English language learners, differences-in-differences

Impact Statement

We used a quasi-experimental design to investigate the effect of the *Bridges in Mathematics* curriculum on student mathematics achievement gains in fifth grade in a large, urban school district. Students who received the curriculum grew measurably more on mathematics scores from the fall to spring of fifth grade than students who received the previous curriculum. English language learners with heterogenous English language proficiency levels and English-proficient peers benefited similarly from the curriculum.

Examining *Bridges in Mathematics* and Differential Effects Among English Language Learners

Promoting students' mastery of basic mathematics facts, concepts, and procedures (National Mathematics Advisory Panel [NMAP], 2008) at scale requires further expansion of evidence-based mathematics curricula that align with the developmental aspects of mathematics learning. A variety of factors speak to this need. Despite increases in fourth graders' mathematics performance on National Assessment of Educational Progress (NAEP) from 1990 to 2019 (Hussar et al., 2020), U.S. fourth graders' mathematics performance in the 2015 Trends in International in Mathematics and Sciences Study (TIMSS) shows they performed significantly behind ten other countries (Hussar et al., 2020). These gaps occur even in light of national policies (e.g., No Child Left Behind [NCLB], 2002; Every Student Succeeds Act [ESSA], 2015) and standards (e.g., implementation of the Common Core State Standards in Mathematics [CCSSM] in 2010; National Governors Association Center for Best Practices [NGA], Council of Chief State School Offers [CCSO], 2010) implemented to improve achievement. Moreover, there have been many efforts to test elementary mathematics intervention programs or strategies, but only two comprehensive elementary mathematics curricula meet What Works Clearinghouse's (WWC) effectiveness definition (*Odyssey Math* and *Everyday Mathematics*; WWC, 2020a; see also Agodini and Harris [2010] who made a similar observation of WWC-reviewed interventions in 2009). Systematically higher effect sizes of curriculum evaluations from developers compared to independent teams may further cloud the evidence base (Wolf et al., 2020).

Elementary school is a key period to ensure development of core mathematics competencies. Acquisition of a variety of core skills is necessary to reach numerous benchmarks in mathematics skill development, especially as students approach middle school. Elementary word-problem solving (Fuchs et al., 2014) as well as mathematical equivalence knowledge

(Matthews & Fuchs, 2020) begin a trajectory towards algebra competence. Additionally, early fraction competencies in elementary school lay a foundation for success in middle school algebra (Booth & Newton, 2012). Koon and Davis (2019) show that grade five mathematics achievement more robustly predicted meeting grade 11 mathematics benchmarks than patterns of students' mathematics courses between grades six and 11. Rigorous core instruction in key areas of mathematical competence remains a critical aspect of effective prevention of mathematics difficulties (Clarke et al., 2011; Clarke et al., 2015).

Multitiered systems of support (MTSS) explicitly embeds prevention within its levels of support (Burns, 2011). Universal curriculum (tier 1 instruction) that all students receive within the general classroom prevents compounding difficulties at broad scale by sufficiently providing the requisite skills for sustained academic development (Mellard et al., 2010). Examining the effectiveness of universal curriculum to ensure students are acquiring core skills is central to effective MTSS and curriculum implementation since effective instruction at tiers 2 and above are premised on adequate universal instruction (Mellard et al., 2010). Research examining universal mathematics curriculum effectiveness among English language learners (ELLs) is in particular need of further development considering the dearth of studies in this area. Equity issues are critical to address in the universal tier in order to ensure all students equitably receive necessary core content (Albers & Martinez, 2015; Robinson-Cimpian et al., 2016).

The current evidence base, achievement trends, and equity issues impacting students' learning necessitate continued examination of practices and programs that support students' learning with added attention to linguistic diversity. Some studies have studied implementation of universal core curricula in randomized trials (e.g., Agodini & Harris, 2010; Clarke et al., 2011; Clarke et al., 2015). Doabler, Clarke et al. (2016) extended Clarke et al. (2011) by testing

the curriculum's effects among Spanish-speaking ELLs, finding positive results of the intervention across baseline performance levels. Yet more work is needed in this area. Additional studies situated in context must examine the causal effects of curricula on student achievement to generate evidence rooted in day-to-day practices. Randomized trials have been a leading method to the evidence-based education movement (Borman, 2009). However, research designs that leverage local decisions to measure causal effects also inform the evidence base of curricula and provide insight into achievement trends among diverse learners in typical educational settings. In our current study, we use the temporal variation in the implementation of a CCSSM-based curriculum in a large school district to investigate treatment effects on annual student-level mathematics achievement gains and effect heterogeneity across ELLs' English language proficiency (ELP) levels.

Standards-Based Instruction in Mathematics

CCSSM implementation in 2010 prompted the need for universal curricula that were well-aligned to these standards, helping to align identification, prevention, and intervention services with benchmarks for students' content mastery. Standards like CCSSM require deliberate, high-fidelity implementation that is well-supported professionally and corresponds with high levels of accountability (Coburn et al., 2016). Disconnects between implementation efforts and accountability to fidelity may generate tensions that weaken teachers' buy-in and standards adherence (Coburn et al., 2016). Practices that align to CCSSM associate with achievement in multiple ways. Schmidt and Huoang (2012) found that CCSSM not only mirrored the standards of higher performing countries on 1995 TIMSS mathematics but also that state standards more similar to CCSSM were associated with higher NAEP mathematics scores

in 2009. Measures of alignment to state-level standards further support small increases in the cognitive demand of CCSSM compared to other standards in 2010 (Porter et al., 2011).

Although a significant majority of the U.S. adopted the CCSSM upon its implementation in 2010, substantial variability in CCSSM implementation remained (Cogan et al., 2013). Cogan et al. surveyed mathematics teachers in 41 U.S. states in 2011 and showed that state standards informed a majority of teachers' instruction across grades 1-12 and that teachers were receptive to CCSSM implementation. Yet, across 40 states, 0% to 35% of teachers relied on textbooks that may not have reflected CCSSM. Variation in materials may impede standards adherence, and teachers indicated PD as well as accessible implementation resources would support CCSSM implementation (Cogan et al., 2013).

Adapting curricula to fit situational needs while retaining standards-alignment and accountability is difficult without the necessary supports that help teachers carry out their instructional strategies (Coburn et al., 2016). Teacher professional development (PD) facilitates high-fidelity implementation and shapes teachers' perceptions of implementing the curriculum, especially in light of standards adherence and assessment-based accountability (McGee et al., 2013). A randomized trial of elementary-level teacher PD (Garet et al., 2016) found positive effects on teachers' mathematics content knowledge and some aspects of instructional quality. Garet et al. found null effects on student achievement, though prior work has provided evidence of a relationship between teachers' mathematics content knowledge and student achievement (Campbell et al., 2014; Hill et al., 2005). Additionally, Blank and de las Alas (2009) established post-test and pre-post effect sizes of 0.13 and 0.21, respectively, of elementary PD in mathematics on student achievement.

These prior studies suggest that improved student achievement likely occurs at the confluence of high-fidelity implementation of standards-aligned curricula, PD for teachers, and access to evidence-based instruction. Decisions about curriculum implementation and, ultimately, the prevention efforts that undergird universal instruction, must stem from clear sources of evidence that reflect day-to-day practices, however. To that end, empirical investigation of curriculum in-context helps determine how policymaker decisions affect student achievement. Uses of quasi-experimental designs (QEDs) are increasing in education research given their ability to assess these exact types of decisions (Gopalan et al., 2020).

English Language Learners and Mathematics Achievement

Mathematics development, prevention of difficulties, and the implementation of standards-based instruction also requires particular attention among ELLs. Multiple sources of language factor into ELLs' academic learning, including general and specific academic language in English (Baker et al., 2014; Doabler, Nelson, et al., 2016) and students' native languages. There are pressing equity issues impacting ELLs' learning in core instruction as well (Robinson-Cimpian et al., 2016). However, mathematics curriculum effectiveness research among ELLs is limited, and more research is needed to better understand how standards-based mathematics curriculum implementation intersects with linguistic diversity.

Students identified as ELLs comprised 10.1% of the student population in the U.S in 2017 (Hussar et al., 2020). Typical reports of ELLs' mathematics performance document significant gaps from their non-ELL peers (e.g., Hussar et al., 2020), though other accounts of linguistically diverse student achievement (using more inclusive definitions of linguistic diversity than ELL) show NAEP mathematics improvements at higher rates relative to monolingual peers (Kieffer & Thompson, 2018). Examining patterns of achievement in relation

to curriculum change informs how new practices and implementation strategies might support ELLs' learning. However, definitions and identification criteria for ELLs varies across states as well as research studies. Research that examines differences in achievement level and growth across ELL status as well as across the ELP continuum is needed to better understand the relationships of language to mathematics and how mathematics curricula support ELLs' learning.

Language holds important relationships to mathematics performance (Vukovic & Lesaux, 2013; Chow & Ekholm, 2019). ELLs in particular face the challenge of acquiring both general and content-specific English language in addition to competencies in specific academic areas (Baker et al., 2014; Doabler, Nelson, et al., 2016). This occurs concurrently with native language development for many students as well. Prior experimental studies have posited that verbalizing problem-solving and meaningful use of content-specific language in tier 1 (Doabler, Clarke, et al., 2016) as well as tier 2 curriculum (Doabler et al., 2019) facilitates ELLs' mathematics development. Vukovic and Lesaux (2013) found that oral language predicted mathematics performance in elementary school, though this relationship was more robust for conceptually-focused mathematics tasks. In another study, elementary students' language syntax skills uniquely positively predicted mathematics performance (Chow & Ekholm, 2019). Mathematics language also relates to mathematics performance among young children (Purpura & Reid, 2016; Purpura et al., 2017). Relationships of vocabulary and reading comprehension may be also informative in this context. Quinn and colleagues (2015) found that growth in vocabulary predicted growth in reading comprehension among mostly English-speaking students, supporting the instrumentalist hypothesis of reading comprehension (Anderson & Freebody, 1981). Relatedly, Lesaux et al. (2010) found that oral English language predicted English reading comprehension among native Spanish-speaking students. Altogether, these findings support that

vocabulary and language acquisition may facilitate comprehension and mathematics performance in a variety of settings and grade levels. For ELLs, building vocabulary and language use may be particularly important for mathematics learning, which is consistent with recommended practice (Baker et al., 2014; Doabler, Nelson, et al., 2016).

However, there is limited research in the influence of mathematics curricula on ELLs' mathematics achievement, and considering ELP level heterogeneity is important for MTSS implementation (Albers & Martinez, 2015). Intervention research within an MTSS model is informative for this area of work, though research on heterogenous effects of mathematics interventions for ELLs remains limited. Doabler, Clarke, et al. (2016) found that a universal kindergarten mathematics curriculum produced similar achievement gains for ELLs across prior achievement levels. Doabler et al.'s (2019) results regarding variation of tier 2 mathematics intervention effects across ELP levels were mixed. Small-scale studies (e.g., Driver & Powell, 2017) or single-case designs (e.g., Orosco, 2014; Leauvano & Collins, 2020) have further investigated the effects of interventions of ELLs at-risk for mathematics difficulties.

These studies also underscore the variability in defining ELL and ELP, necessitating further examination of how mathematics curriculum and ELP relate to achievement. Studies also vary in the scale and scope of intervention and generally target a more specific subgroup of students (e.g., ELLs at-risk for mathematics difficulties). QEDs that leverage day-to-day decision making of policymakers to identify causal effects of mathematics curriculum offer a particularly important perspective in this literature as they can provide insight into how implementing new practices at wide scale supports ELLs' learning.

The Present Study

In the present study, we investigated the effects of the implementation of a standards-based curriculum, *Bridges in Mathematics* (The Math Learning Center, 2015), on fifth graders' annual mathematics achievement gains within a large, urban, midwestern school district compared to the district's former curriculum (*Investigations*). We also examined the extent to which these effects varied across ELP levels.

Bridges is implemented internationally and has received strong reviews from curriculum reviewers (e.g., Education Reports, 2018). It emphasizes instruction that is "[…] linguistically, visually, and kinesthetically rich […]" (The Math Learning Center, 2020, n.p.) and is designed for 80 minutes of mathematics instruction per day (The Math Learning Center, n.d.a.). *Bridges*'s publisher provides some information of the curriculum's research base (The Math Learning Center, n.d.b.), yet we are aware of only one independent causal analysis of this curriculum's impact on student achievement (SEG Measurement, 2018).

Unlike *Bridges*, which was designed to align directly to CCSSM standards, *Investigations* required adaptations by the district to meet needs for CCSSM alignment. At the time of this study, some schools had implemented *Bridges in Mathematics* curriculum and some schools were still using *Investigations*. *Investigations* had been used in the district for approximately 10 years. This version of Investigations was not aligned to the CCSSM and therefore lacked explicit emphasis on core aspects of CCSSM. Key to the *Bridges* curriculum is the nonlinear sequence of CCSSM coverage that builds a continuous cycle of scaffolding for content mastery (Hansen, 2017). This cyclical pattern of content coverage – e.g., introducing a standard in small increments and continuously revisiting the standard to promote mastery – constitutes "meaningful distributed practice" (MDP; Hansen, 2017) and parallels the sequencing of standards ("spiraling") in a different curriculum meeting WWC's effectiveness criterion –

Everyday Mathematics (Center for Elementary Mathematics and Science Education, n.d.). However, this cycle also imposes potential implementation barriers given that it was not the typical linear and sequential sequence of standards coverage that was central to *Investigations*, which provided fewer distributed practice opportunities on the same standards over time. A number of reviews (Dunlosky et al., 2013; Gerbier & Toppino, 2015; Son & Simon, 2012) highlight the promising positive effects of distributed compared to massed practice opportunities. *Investigations'* standards coverage structure was more massed in nature.

Although the curricula naturally differed along important lines such as CCSSM-aligned content and the structure of standards coverage, the clearest distinction between the two curricula in this district was the PD and implementation support that teachers received to implement *Bridges*, particularly with respect to the standards coverage. This resulted in a greater emphasis on instruction on the specific CCSSM elements of *Bridges* and the cycling of standards, whereas *Investigations* lacked the built-in alignment to CCSSM as well as explicit instructional support guiding CCSSM implementation. The prior curriculum may have had content emphases on language and visuals to a similar degree of *Bridges* (in addition to individual teacher instructional strategies not built-in to the curriculum as well as modifications to better align to CCSSM), but the district clearly differentiated the two curricula with PD and implementation support for *Bridges*-specific elements (e.g., content coverage, lesson planning, instructional strategies aligned to the curriculum).

Our control condition in this case parallels what Cogan et al. (2013) echoed in their survey of teachers' use of curriculum materials relate to the CCSSM. Students in schools implementing *Bridges* received instruction under content explicitly aligned to CCSSM from teachers that received PD targeted to promote instruction that balanced instructional flexibility

with high implementation fidelity. Specific PD and implementation supports ideally drew more attention to the unique elements of *Bridges* and the CCSSM standards guiding learning that may not have been explicit within the curriculum or instructional strategies in prior practice.

We investigated the main effects of the *Bridges* curriculum on fifth graders' mathematics achievement gains and whether this effect varied by ELP level. We addressed two research questions:

1. What is the impact of *Bridges in Mathematics* on students' yearly mathematics achievement gains in fifth grade?

We did not express a strong a priori hypothesis regarding the impact of *Bridges* on student achievement. We suspected that if the curriculum potentially improved student achievement gains, observed effects would likely operate through PD, implementation support, and standards alignment of *Bridges* coupled with instruction and the curricular content. However, new implementation and PD could have also led to decreases in fidelity that did not improve implementation beyond business-as-usual practices. We believed the more likely scenario entailed strong PD, greater implementation fidelity, and improved instruction.

2. Is the effect of *Bridges* different across levels of ELP?

Bridges' focus on using mathematics language potentially provided enhanced opportunities for ELLs to use language in their mathematics learning (e.g., vocabulary). Prior work has noted the importance of language use in ELLs' mathematics learning (Doabler, Clarke, 2016; Doabler, Nelson, et al., 2016; Doabler et al., 2019), and oral English language predicts reading comprehension (Lesaux et al., 2010) as well as mathematics performance (Vukovic & Lesaux, 2013). Although the majority of classroom teachers did not receive *Bridges* PD that explicitly

targeted teaching ELLs, we believed the curriculum structure, content, and increased fidelity had potential to be differentially beneficial for ELLs with varying ELP levels.

Method

Procedure

We used student-level data from fifth-grade students in a large, urban, midwestern school district (Madison Metropolitan School District [MMSD]) to study the effects of *Bridges* curriculum implementation on student achievement growth and differential effects across ELP levels. Procedures for this study were reviewed and approved as exempt by Florida State University Institutional Review Board (IRB; Study 00001629)

Student Sample

We used data from fifth-grade students in 29 of MMSD's 32 elementary schools (three schools enroll only kindergarten to second-grade students). The district implemented *Bridges* in three phases of schools across three years (phase one [2016-17], two [2017-18], and three [2018-19]). Out of the 29 schools, phase one (nine schools) and phase two (12 schools) had completed implementation at the time of this study. The staggered implementation of the curriculum facilitated identifying control group students from concurrent non-*Bridges* schools as well students in the district prior to any *Bridges* implementation. To do this, we used data from three year-based cohorts: fifth graders in 2015-16, 2016-17, and 2017-18. Students in the control group comprised students in the district prior to any *Bridges* implementation (i.e., 2015-16) as well as students in schools who had not yet received *Bridges* instruction (i.e., non-*Bridges* students in 2016-17 and 2017-18). Table 1 displays the timing of the student cohorts and phases of *Bridges* implementation between grades three and five.

<Table 1 about here>

We used students' exposure to *Bridges* in fifth grade to analyze the effect of *Bridges* on mathematics achievement gains (this includes the fifth-grade data of students who also received *Bridges* in fourth grade). The full sample of students with a valid school identification code in grade five, grade five demographic data, and an ELP level (from the prior year) totaled 5,555 students. No demographic data were missing among these 5,555 students except for a proportion of parent education level (which was the highest- or only-reported level between grades three and five); we generated an indicator variable to capture remaining missingness. We exclude students who were missing fall of grade five reading assessment scores (3.5%), so our sample of students with complete covariate data totaled 5,359. Missing values on our outcome (3.1%) further limited our final sample to 5,193 fifth-grade students in MMSD from 2015-16 (n = 1,765), 2016-17 (n =1,695), and 2017-18 (n = 1,733). Treatment group students (n = 1,839) were those who received *Bridges* instruction in 2016-17 (phase one implementation) or 2017-18 (phase two implementation). The remaining 3,354 students in 2015-16, 2016-17, and 2017-18 comprised the control group.

Measures

Dependent variable

Measures of Academic Progress Math Change Score. We calculate grade five change on Northwest Evaluation Association's Measures of Academic Progress (MAP) math assessment by subtracting fall MAP achievement scores from spring scores in the same year. MAP math is computer-adaptive, vertically scaled in Rasch (RIT) units, and shows adequate reliability in the normative sample: fall-spring test-retest reliability is .90 and marginal reliability is .97 in fall and spring (National Center on Intensive Intervention [NCII], 2019).

Independent variables

Bridges Exposure. We have available a single school identification variable for fifth graders reflecting the last school students attended during fifth grade. We use this indicator to determine *Bridges* exposure. Students indicated to be enrolled in a *Bridges* school in fifth grade were given a 1; all other students were given a 0. As Table 1 shows, some students would have received *Bridges* in fourth grade as well. Students that switched from a *Bridges* school to a non-*Bridges* school between fourth and fifth grade (only a small percent) were considered part of the control group. We do not capture students who may have switched schools during the year of analysis. However, most students (94%) remained in the same school between fourth and fifth grade based on school indicators from both years.

As aforementioned, *Bridges* implementation occurred in a three-phase rollout. To support *Bridges* implementation, most teachers participated in a two-day workshop in the summer prior to the first year of implementation at their school, which was presented by *Bridges* trainers from The Math Learning Center. At the workshop, teachers learned about the key components of the *Bridges* curriculum (unit structure and lesson components) and specific *Bridges* instructional strategies. Teachers were grouped into grade bands (kindergarten to second grade and third to fifth grade) during the PD in order to cater to unique grade-specific implementation factors. Beginning in the fall with the start of the school year, district-level curriculum and instruction staff met monthly with grade-level teacher teams in schools for the first year of implementation. The focus for monthly work was to check-in on implementation and provide support for teacher collaboration with a preview of upcoming units and lessons for the grade level. Monthly grade-level meetings with teachers and building leadership provided implementation support to increase the instructional flexibility needed to creatively intermix standards coverage while also promoting fidelity to the curriculum and CCSSM.

ELP. Students in this state and district are assigned an ELP level from 1 to 7. ELLs who have attained English proficiency are assigned a 6, levels 1-5 reflect increasing degrees of proficiency, and level 7 reflects native English speakers (Wisconsin Department of Public Instruction, n.d.a). ELLs' ELP level is based primarily on scores from the ELP assessment used across the state (Assessing Comprehension and Communication in English State-to-State for ELLs [ACCESS] from World Class Instructional Design and Assessment [WIDA]; Wisconsin Department of Public Instruction, n.d.a). ACCESS is administered in the winter of each school year, and it has strong validity and reliability (Wisconsin Department of Public Instruction, n.d.b). Proficiency scores range from 1.0 to 6.0 (rounded to the nearest tenth), and the scale (1.0 to 6.0) remains the same regardless of students' grade. ACCESS scores are truncated to the next whole number to create categorical ELP levels (e.g., 3.0 – 3.9 truncates to ELP 3). Proficiency scores from students' immediate previous grade informs their ELP level in the subsequent grade. However, in our data, ELP level does not reflect exactly who would have been limited in ELP in grade five. ACCESS scores from the prior year between 5 and 5.9 equated to an ELP of 5, which did not correspond to local district ACCESS proficiency criteria (in general, ELP levels were left as-reported even if they did not align to ACCESS scores). Also, changes to ACCESS in 2016-17 potentially resulted in lower proficiency scores and altered proficiency criteria after 2016-17 (Wisconsin Department of Public Instruction, n.d.c). Some students with an ELP level of 6 or below did not have ACCESS scores (14.3%). This could be a result of a variety of factors, though it is likely due to prior reclassification (i.e., they no longer took ACCESS). We include only students enrolled in MMSD public schools in the immediate prior year that had an ELP level (or an ACCESS score for analyses limited to students who took ACCESS).

We combined ELP levels (levels 6-7 = proficient, 4-5 = "mid" ELP, 1-3= "low" ELP) so we could measure effect heterogeneity across levels of ELP in the full sample. There are limitations to this method, however. We do not capture variation in the length of time since reaching English proficiency (i.e., those with ELP of 6); we were primarily concerned with identifying the level of proficiency relevant to students' fifth grade year. Our labels of "mid" and "low" ELP likely do not reflect the complexity of skills that each ELP level intends to represent. Our "low" ELP definition is more representative of students with ELP of 3 since students with ELP of 2 or below typically do not take MAP in MMSD (though this is not the case for all students). "Mid" and "low" designations should be interpreted in relative terms as opposed to absolute levels of proficiency. We also used ACCESS Overall Composite Score proficiency levels (from the prior year) to address research question two.

Covariates. Covariates in our analysis include indicators for parent education level, free/reduced-price lunch eligibility, race/ethnicity, gender (male or female), individualized education plan (IEP) status (yes or no IEP), student cohort year (2015-16, 2016-17, 2017-18), and fall of grade five MAP reading achievement. All demographic covariates were measured during fifth grade except parent education, for which we used the highest- or only-reported level across grades three to five (missing parent education level was given a new indicator variable). We also included indicators for assignment to implementation cohorts of a one-to-one student technology plan, which began implementation in 2015-16. The technology plan cohorts to which schools were assigned (six in total) may have related to schools' decision to implement *Bridges*. We assigned technology cohort indicators to schools regardless of schools' year-based cohort in this study (i.e., if a school was assigned to implement technology in 2015-16, we assigned the

corresponding technology implementation cohort indicator across all years in this study rather than only in the year the technology was implemented).

Treatment Balance

Demographic and ELP characteristics of our analytic sample along with treatment balance tests are provided in Table 2.

<Table 2 about here>

Treatment groups show balance across the majority of characteristics except for some differences across parent education and ELP level. ELP differences across groups may relate to changes in the ELP measure in 2016-17.

Table 3 displays fall and spring MAP achievement data as well as prior year ELP scores for *Bridges* and comparison students.

<Table 3 about here>

Data in Table 3 indicates that treatment conditions achieved similarly on fall grade five MAP. ACCESS score differences across conditions are not significant when controlling for cohort year.

Design

We used a pre-post differences-in-differences (DiD) design (Lechner, 2011) to test the effect of *Bridges* on student mathematics achievement change scores from the fall to spring of fifth grade and variation in the treatment effect across ELP. We identify the model based on temporal variation in the exposure of three different cohorts of fifth graders to *Bridges*.

Missing data

Among students with complete covariate data (n = 5,359), 3.1% were missing both fall and spring MAP math scores, limiting the analytic sample to 5,193 students. The majority of MAP math scores were missing from spring. Proportions of missing MAP change scores did not

vary by condition (*Bridges* = .029, *Investigations* = .032, $M_{\text{Diff}}(SE)$= .003(0.005), p = .55). We used inverse probability weighting (IPW; Seaman & White, 2013) to reduce potential bias induced by missing data. See supplemental materials for more information on our IPW method.

Analysis

We use a single-level ordinary-least squares (OLS) regression model to estimate the main effect of *Bridges* on students' fall-spring of fifth grade change scores. Change scores have valuable properties for estimating causal effects in DiD designs (Kim & Steiner, 2019) and are similar to a student fixed effect that removes the association between time-invariant attributes of students and their academic achievement. We also use this model to assess variation across three-category ELP level (proficient, mid, low). We use cluster-robust standard errors to correct errors for the nonindependence of students nested within schools. We do not use multilevel models as our primary analysis, though we use *i* to indicate student-level variables and *j* to indicate school-level variables (robustness checks of our results include multilevel models, however). Our primary model, Model 1, is as follows:

$$\begin{aligned} \Delta MAP\ Math_{ij} = \beta_0 + \beta_1\ Bridges_j + \beta_2\ ELP\ Level_{ij} + \\ \beta_3\ ELP\ Level_{ij}\ \times Bridges_j + \\ \beta_4\ Cohort\ 2_{ij}\ \times Bridges_j + \beta X_{ij} + \varepsilon_{ij} \end{aligned} \quad (1)$$

where $\Delta MAP\ Math_{ij}$ is the annual change score outcome for student *i* in school *j*, $Bridges_j$ is an indicator for the schools implementing Bridges in a given year, and $ELP\ Level_{ij}$ represents indicators of students' three-category ELP level (proficient as reference group). $\beta_3 ELP\ Level_{ij}\ \times Bridges_j$ is the interaction between Bridges exposure and ELP level (this separates into two interaction terms due to two indicators variables for the three ELP levels). Because we are interested in the average effect of Bridges regardless of when it was implemented (i.e., 2016-17 or 2017-18), $\beta_4\ Cohort\ 2_{ij}\ \times Bridges_j$ interacts the cohort 2 and

Bridges indicator. This controls for variation of the effect of Bridges across year cohorts. X_{ij} represents a vector of covariates that includes fall grade five MAP reading performance and dummy-codes for demographic variables displayed in Table 2 as well as student cohort year (2017 as reference group). We also include five dummy codes for the school cohorts (six total) of one-to-one technology implementation (first cohort [2015-16] is the reference group). ε_{ij} is a random error term adjusted for clustering at the school by cohort level (resulting in 87 clusters, one for each school within each year cohort; see supplemental materials for more on this).

We also use prior year ACCESS Overall Composite Score proficiency levels to test the variation of the effect of Bridges across ELP. We use a single-level OLS model (Model 2) for this analysis

$$\begin{aligned} \Delta MAP\ Math_{ij} = \beta_0 + \beta_{01}\ Bridges_j + \beta_2\ ACCESS\ Composite_{ij} + \\ \beta_3\ ACCESS\ Composite_{ij} \times Bridges_j + \\ \beta_4\ Cohort\ 2_{ij} \times Bridges_j + \beta X_{ij} + \varepsilon_{ij} \end{aligned} \quad (2)$$

Besides the variable used for terms β_2 and β_3, this model mirrors Model 1, and our estimands of interest are again β_1 and β_3. We use the same IPWs for Models 1 and 2, though Model 2 is restricted to only the students with available ACCESS overall composite proficiency scores from the prior year (n = 1,367). ε_{ij} is also corrected for clustering at the school by cohort level.

Results

To investigate the effect of *Bridges* and the interaction terms with ELP level in Models 1 and 2, we use Stata's (StataCorp, 2019) *margins* command to calculate average marginal effects (AMEs; see Williams, 2012) assuming unbalanced groups for the coefficients of interest (β_1 and β_3) in Models 1 and 2. AMEs represent the difference in the outcome per one-unit change in the predictor averaged across every individual's values on the covariates (and interactions) in the model (Williams, 2012). We use change score standard deviations to estimate Hedges's *g* effect

sizes (Hedges, 1981) using the independent groups formula described in WWC (2020b). We use Stata's *esizei* command to calculate Hedges's *g*.

In online supplemental materials, Table S1 displays unadjusted change scores, adjusted change scores, AMEs, and effect size estimates for research questions 1 and 2 in change score *SD*s. In Supplemental Table S2, we present the Benjamini-Hochberg (BH) correction for controlling for the false-discovery rate in our main hypothesis tests.

Research Question 1

For research question 1, we attend to the predicted difference in change scores between our treatment and comparison groups averaged across all covariate terms (including interactions) in Model 1 (i.e., the AME of β_{01} in Model 1). Students receiving *Bridges* gained significantly more on MAP math than comparison students (b = 2.022, SE = 0.616, t = 3.28, p = .001). Adjusting for covariates, *Bridges* students gained an average of 11.24 points on MAP math from fall to spring; comparison students gained 9.21 points. This main effect of 2.022 equates to a Hedges's *g* of 0.250 (standardized on change score standard deviations). Figure 1 displays the predicted change scores for each treatment group from Model 1 (top panel) and the estimated difference in change scores between conditions (AMEs; bottom panel). Supplemental Table S3 displays full regression results of Model 1 (however, these are not AMEs and do not represent tests of our main effect and interactions with ELP).

Research Question 2

Figure 1 also displays the simple effects of *Bridges* estimated at each ELP level (i.e., using the *at* instead of the *over* command in Stata's *margins*). These effects translate to effect sizes (*g*) of 0.226 for English language-proficient students (ELP levels 6-7), 0.336 for students with mid ELP (ELP levels 4-5), and 0.337 for students with low ELP (ELP levels 1-3; *SD*s for

each treatment group within ELP levels can be found in Supplemental Table S1). We conducted four pairwise comparisons of the ELP level simple effects to test whether ELP moderates *Bridges.* None of these tests were statistically significant. The effect of *Bridges* was not significantly greater among mid ELP levels compared to proficient (b = 0.617, SE = 0.839, t = 0.74, p = .464) or low ELP levels (b = 1.063, SE = 1.072, t = 0.99, p = .324). A reverse-Helmert contrast scheme comparing the proficient category to combined mid and low ELP (contrast terms of -1, .631, and .369, respectively) was not significant *Bridges* (b = 0.782, SE = 0.708, t = 1.10, p = .273). Last, the difference between mid and low ELP levels was not significant (b = 0.445, SE = 1.255, t = 0.35, p = .724). We conclude the effect of *Bridges* does not vary significantly across ELP levels.

<Figure 1 about here>

Results from Model 2 show the effect of *Bridges* does not significantly vary across ACCESS performance. Figure 2 displays the AME of *Bridges* at specific values of ACCESS. The total difference in the AME of *Bridges* across ACCESS scores 3 to 6 was small in magnitude and not statistically significant (b = -0.101, SE = 2.325, t = 0.04, p = .965). The main effect of *Bridges* remains significant in Model 2 (b = 2.935, SE = 0.964, t = 3.05, p = .003, g = 0.376). Supplemental Table S4 displays full results of Model 2.

<Figure 2 about here>

Robustness Checks

We conducted robustness checks of our DiD estimates to assess robustness of our estimation model, the internal validity of our DiD design, and the sensitivity of our results to omitted variables. We provide a brief overview here; see the “Robustness Checks” section for additional detail. Generally, our results were robust to different model specifications, outliers, the

length of time students were exposed to *Bridges* (i.e., some students already received *Bridges* in fourth grade), whether students switched schools between the prior and current year of analysis, and prior year IEP status (since current status could change during the year). Analyses using current year (grade five) ELP/ACCESS scores display a pattern of results similar to analyses with prior year ELP/ACCESS. Overall, robustness tests do not suggest significant internal validity threats to our DiD design, though these tests do not rule out all possible threats. A sensitivity analysis indicates that an omitted variable would have to partially correlate at least .137 with both *Bridges* exposure and MAP mathematics change score (impact = .019 [i.e., .137*.137]) to change our main estimate (b = 2.022, SE = 0.616) to a null effect (Frank, 2000).

Discussion

We used the staggered implementation of a standards-based elementary mathematics curriculum in a large school district to estimate the causal effect of *Bridges* on annual mathematics change scores. On average, students in *Bridges* schools gained approximately two more points on MAP from fall to spring than students who received the previous curriculum, *Investigations* (g = 0.250). Universal mathematics curriculum and instruction are key sources of prevention within MTSS (Clarke et al., 2011; Clarke et al., 2015, Mellard et al., 2010), and more research is needed in this area, particularly studies focusing on ELLs (e.g., Doabler, Clarke, et al., 2016). This implementation of *Bridges* offered a unique opportunity to examine effects of universal curriculum changes on student achievement and effect heterogeneity among ELLs. QEDs are growing in popularity in education research (Gopalan et al., 2020). Our specific QED, DiD, leverages policy changes or exposures that occur in day-to-day educational settings that can be highly informative of how policies and practices impact student achievement over time. QEDs

are a critical tool for education researchers and practitioners to test causal effects of systematic changes to practice, which can inform effectiveness both locally and potentially more broadly.

The primary implication of our work is that *Bridges*, a widely used, standards-based curriculum, may possess meaningful advantage to students beyond the previous curriculum implemented in the district. These observed effects may be a combination of curricular content, PD, and implementation support. Prior work has shown that PD is effective in bolstering aspects of teacher knowledge and instruction (Garet et al., 2016), and meta-analyses show it may meaningfully improve student achievement (Blank & de Alas, 2009). Coupling the PD for *Bridges* was ongoing implementation support for schools in their first year of *Bridges* implementation. As aforementioned, implementation support was in part focused on the sequencing of standards and lessons throughout the curriculum that differed significantly from the prior curriculum. Our QED is an important step in building a stronger evidence base for universal mathematics curricula by examining these locally-driven, district-wide efforts to change implementation and instruction. QEDs have a meaningful role in evidence-based education by testing causal questions in a manner that is timely, feasible, and rooted in realistic conditions of instruction, such as systematic changes to core curriculum. Our work extends that of prior randomized trials of universal mathematics curricula (e.g., Agodini & Harris, 2010; Clarke et al., 2011; Clarke et al., 2015) and the effects among ELLs (Doabler, Clarke, et al., 2016).

Bridges implementation related to achievement gains similarly across ELP levels. Prior work has shown effective teachers are effective for non-ELLs and ELLs alike (Loeb et al., 2014). These effects on change scores should also be taken in light of the pre-existing differences in mathematics achievement levels: students with mid and low ELP began fifth grade performing

0.56 and 1.23 *SD*s (*g*), respectively, below English proficient peers. Identifying the mechanisms of learning in core instruction that bolster the achievement trajectories of linguistically diverse students is needed to better understand universal curriculum implementation and how to continually support equity for ELLs in core instruction (Robinson-Cimpian et al., 2016). For example, explicit instruction may afford ELLs the key opportunities for developing early mathematics skills through deliberate use of language in problem solving (Doabler, Clarke, 2016; Doabler, Nelson, et al., 2016; Doabler et al., 2019). Both general language (Chow & Ekholm, 2019; Vukovic & Lesaux, 2013) and specific mathematics language (Purpura & Reid, 2016; Purpura et al., 2017) predict mathematics performance, and oral language (Lesaux et al., 2010) and vocabulary (Quinn et al., 2015) predict reading comprehension. Yet the lack of differential effects of *Bridges* across ELP levels requires additional attention to understand how ELP intersects with instruction (Albers & Martinez, 2015) and how language facilitates mathematics learning.

Limitations and Future Directions

Other sources of selection bias into curriculum implementation may remain in our analysis. For example, systematic variation in the readiness or capacity of schools to implement with fidelity will be important to consider in future work. These factors could relate to achievement gains and may not be accounted for in our set of covariates, so these should be more explicitly measured in future studies that do not randomize treatment assignment. Additionally, the current estimates do not reflect within-year entry or exit into a school using the curriculum. Differences in the ELP assessment across years remains a substantial limitation, though controlling for cohort year helps address this.

We did not have data that would inform curriculum mechanisms, such as fidelity, dosage, mathematics subskills, or other mediators of intervention effects. Future research should attend to these issues as part of program evaluation efforts in order to target key competencies at critical junctures of mathematics development (e.g., fraction knowledge prior to middle school [Booth & Newton, 2012]). This could also help promote system-wide efforts to support universal instruction by identifying key elements of fidelity and instruction like instructional quality and teacher mathematics knowledge (Garet et al., 2016). Classroom observations of core instruction mechanisms would also be highly informative, and prior studies have included this aspect (Doabler, Clarke, et al., 2016). Relatedly, identifying those who were actually "treated" or received the core mechanisms of the curriculum will be important to identify in future research (e.g., Schochet & Chiang, 2011). Long-term intervention impacts are also important to investigate (Bailey et al., 2020). Our limited scope of analysis should prompt future studies to consider how universal curriculum implementation can serve its intended role in long-term prevention.

It is also important to consider that our outcome was an English measure of mathematics and how this may impact assessment of ELLs' mathematics knowledge (Solano-Flores, 2016). Studies involving ELLs that assess multiple areas of mathematics skills (e.g., Doabler, Clarke, et al. 2016; Doabler et al., 2019) likely capture more granularity in mathematics skills than what we capture on MAP. Measures of specific mathematics subdomains (including those in native languages) might speak more to how a mathematics curriculum promotes learning. Some existing work has addressed the intersection of specific domains of mathematics performance and specific instruction techniques (explicit instruction) with ELLs in tier 1 (Doabler, Clarke, et al., 2016) and across ELLs' ELP levels in tier 2 intervention (Doabler et al., 2019). More work in

this area is needed, however. Examining these factors may help better target prevention strategies that sustain mathematics development over a longer term in light of the many dimensions of language, mathematics, and implementation challenges.

Conclusion

Bridges, a widely used standards-based curriculum, has received little empirical scrutiny. School districts and policymakers are pressed to make critical decisions about curriculum materials and how to support teachers' curriculum implementation in ways that meets students' needs. Although many possible routes to the effects identified here are possible (e.g., teacher PD and implementation support, more rigorous and standards-aligned content), the empirical evidence we present lends credit to the use of *Bridges* as a comprehensive, standards-based universal curriculum to bolster students' mathematics gains in late elementary school.

References

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary math curricula. *Journal of Research on Educational Effectiveness*, *3*(3), 199-253.

Albers, C. A., & Martinez, R. S. (2015). *Promoting academic success with English Language Learners: Best practices for RTI*. Guildford.

Anderson, R., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). International Reading Association.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, *21*(2), 55-97.

Baker, S., Lesaux, N., Jayanthi, M., Dimino, J., Proctor, C. P., Morris, J., Gersten, R., Dimino, J., Jayanthi, M., Haymond, K., & Newman-Gonchar, R (2014). *Teaching academic content and literacy to English learners in elementary and middle school* (NCEE 2014-4012). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Blank, R. K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement.* Council of Chief State School Officers.

Booth, J. L., & Newton, K. J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology*, *37*(4), 247-253.

Borman, G.D. (2009). The use of randomized trials to inform education policy. In G.Sykes, B. Schneider, D.N.Plank (Eds.), *Handbook of education policy research* (pp. 129–138). Routledge.

Burns, M. K. (2011). School psychology research: Combining ecological theory and prevention science. *School Psychology Review*, *40*(1), 132-139.

Campbell, P. F., Rust, A. H., Nishio, M., DePiper, J. N., Smith, T. M., Frank, T. J., Clark, L. M…Choi, Y. (2014). The relationship between teachers' mathematical content, pedagogical knowledge, teachers' perceptions, and student achievement. *Journal for Research in Mathematics Education*, *45*(4), 419-459. doi: 10.5951/jresmatheduc.45.0419

Center for Elementary Mathematics and Science Education. (n.d.). *The spiral: Why Everyday Mathematics distributes learning.* Author. Retrieved from *https://everydaymath.uchicago.edu/about/research-results/EM_Spiral_20121125.pdf*

Chow, J. C., & Ekholm, E. (2019). Language domains differentially predict mathematics performance in young children. *Early Childhood Research Quarterly*, *46*, 179–186.

Clarke, B., Smolkowski, K., Baker, S. K., Fien, H., Doabler, C. T., & Chard, D. J. (2011). The impact of a comprehensive tier 1 core kindergarten program on the achievement of students at risk in mathematics. *The Elementary School Journal*, *111*(4), 561-584.

Clarke, B., Baker, S. K., Smolkowski, K., Doabler, C., Cary, M. S., & Fien, H. (2015). Investigating the efficacy of a core kindergarten mathematics curriculum to improve student mathematics learning outcomes. *Journal of Research on Educational Effectiveness*, *8*, 303-324.

Coburn, C. E., Hill, H.C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The common core state standards and implementation research. *Educational Researcher, 45*(4)*,* 243-251.

Cogan, L., Schmidt, W., & Huoang, R. (2013). *Implementing the Common Core State Standards for Mathematics: What we know about teachers of mathematics in 41 states*. The

Education Policy Center Working Paper #33. https://files.eric.ed.gov/fulltext/ED558137.pdf

Doabler, C. T., Clarke, B., Kosty, D. B., Baker, S. K., Smolkowski, K., & Fien, H. (2016). Effects of a core kindergarten mathematics curriculum on the mathematics achievement of Spanish-speaking English language learners. *School Psychology Review, 45*(3), 343–361.

Doabler, C. T., Clarke, B., Kosty, D., Smolkowski, K., Nelson, E., Fien, H., & Baker, S. K. (2019). Building number sense among English learners: A multisite randomized controlled trial of a tier 2 kindergarten mathematics intervention. *Early Childhood Research Quarterly*, *47*(2), 432–444.

Doabler, C. T., Nelson, N. J., & Clarke, B. (2016). Adapting evidence-based practices to meet the needs of English learners with mathematics difficulties. *Teaching Exceptional Children, 48*(6), 301-310.

Driver, M. K., & Powell, S. R. (2017). Culturally and linguistically responsive schema intervention: Improving word problem solving for English language learners with mathematics difficulty. *Learning Disability Quarterly*, *40*(1), 41-53.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/ 10.1177/1529100612453266

Education Reports. (2018). *Bridges in Mathematics (2015)*. https://www.edreports.org/reports/overview/bridges-in-mathematics-2015

Every Student Succeeds Act of 2015, 20 U.S.C 70. § 6301 *et seq* (2015).

Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research*, *29*(2), 147-194.

Fuchs, L. S., Powell, S. R., Cirino, P. T., Schumacher, R. F., Marrin, S., Hamlett, C. L…Changas, P. C. (2014). Does calculation or word-problem instruction provide a stronger route to prealgebraic knowledge? *Journal of Educational Psychology, 106*(4), 990–1006.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & Borman, G. D. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher Professional Development* (NCEE 2016-4010). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience and Education*, *4*, 49-59.

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, 218-243.

Hansen, P. (2017). Bridges & meaningful distributive practice. Blog post. Retrieved from: https://www.mathlearningcenter.org/new/blog/bridges-meaningful-distributive-practice

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107-128.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371-406. doi: 10.3102/00028312042002371

Hussar, B., Zhang, J., Hein, S., Wang, X., Roberts, A., Cui, J…Barmer, A. (2020). *The condition of education 2020* (NCES 2020-144). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2020/2020144.pdf

Kieffer, M. J., & Thompson, K. D. (2018). Hidden progress of multilingual students on NAEP. *Educational Researcher*, *47*(6), 391–398.

Kim, Y, & Steiner, P. M. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods and Research*. Advanced online publication.

Koon, S., & Davis, M. (2019). *Math course sequences in grades 6–11 and math achievement in Mississippi* (REL 2019-007). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Lechner, M. (2010). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165-224. https://doi.org/10.1561/0800000014

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, *31*(6), 475-483.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, *36*(4), 457–475.

Luevano, C., & Collins, T. A. (2020). Culturally appropriate math problem-solving instruction with English language learners. *School Psychology Review*, *49*(2), 144-160.

The Math Learning Center. (2015). *Bridges in Mathematics*. Author.

The Math Learning Center. (2020). *Building mathematical thinkers.*

https://www.mathlearningcenter.org/bridges

The Math Learning Center. (n.d.a). *What you need to know about Bridges [Handout]*. https://www.mathlearningcenter.org/sites/default/files/documents/BridgesAssumptions.pdf

The Math Learning Center. (n.d.b.). *Research base for Bridges in Mathematics Second Edition [Handout]*.https://www.mathlearningcenter.org/sites/default/files/documents/Bridges2ResearchBase.pdf

Matthews, P. M., & Fuchs, L. S. (2020). Keys to the gate? Equal sign knowledge at second grade predicts fourth-grade algebra competence. *Child Development*, *91*(1), 14-28.

McGee, J.R., Polly, D., & Wang, C. (2013). Guiding teachers in the use of a standards-based mathematics curriculum: Teacher perceptions and subsequent instructional practices after an intensive PD program. *School Science and Mathematics*, *113*(1), 16-28.

Mellard, D., McKnight, M., Jordan, J. (2010). RTI tier structures and instructional intensity. *Learning Disabilities Practice*, *25*(4), 217-225.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. United States Department of Education.

National Governors Association Center for Best Practices, Council of Chief State School Offers. (2010). *Common core state standards-mathematics*. Author.

National Center on Intensive Intervention (2019). *MAP growth mathematics*. https://charts.intensiveintervention.org/screening/tool/?id=576cd73956493b98

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq* (2002).

Orosco, M. J. (2014). A math intervention for third grade Latino English language learners at

risk for math disabilities. *Exceptionality*, *22*(4), 205-225.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core standards: The new U.S. intended curriculum. *Educational Researcher*, *40*(3), 103-116.

Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, *36*(3)*,* 259-268.

Purpura, D. J., Napoli, A. R., Wehrspann, E. A., & Gold, Z. S. (2017). Causal connections between mathematical language and mathematical knowledge: A dialogic reading intervention. *Journal of Research on Educational Effectiveness, 10*(1), 116-137.

Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relationships between vocabulary knowledge and reading comprehension. A latent change score modelling study. *Child Development*, *86*(1), 159-175.

Robinson-Cimpian, J. P., Thompson, K. D., & Umansky, I. M. (2016). Research and policy considerations for English learner equity. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 129–137.

Schmidt, W.H., & Houang, R.T. (2012). Curricular coherence and the common core state standards for mathematics. *Educational Researcher*, *41*(8), 294-308.

Schochet, P. Z., & Chiang, H. S. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics*, *36*(3), 307-345.

Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3), 278-295.

SEG Measurement. (2018). *An evaluation of the effectiveness of Bridges in*

Mathematics for developing student math skills. Author. https://www.mathlearningcenter.org/sites/default/files/documents/Bridges%20in%20Mathematics%20Effectiveness%20Study.pdf

Solano-Flores, G. (2016). *Assessing English Language Learners*. Routledge.

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review*, 24(3), 379–399. https://doi.org/10.1007/s10648-012-9206-y

StataCorp. (2019). *Stata 16.1 SE* [Computer software]. Author.

Wisconsin Department of Public Instruction. (n.d.a.). ACCESS for ELLs data and reporting. https://dpi.wi.gov/assessment/ELL/data

Wisconsin Department of Public Instruction. (n.d.b.). ACCESS for ELLs . https://dpi.wi.gov/assessment/ell

Wisconsin Department of Public Instruction. (n.d.c.). WISEdash (for districts) ACCESS dashboards. https://dpi.wi.gov/wisedash/districts/about-data/access

Vukovic, R. K., & Lesaux, N. K. (2013). The language of mathematics: Investigating the ways language counts for children's mathematical development. *Journal of Experimental Child Psychology*, *115*(2), 227-244.

What Works Clearinghouse. (2020a). Find what works: Mathematics. Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/FWW/Results?filters=,Math&customFilters=K,1,2,3,4,5,Curriculum, on December, 6th 2020.

What Works Clearinghouse. (2020b). What Works Clearinghouse procedures handbook, version

4.1. Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. 2016.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, *12*(2), 308-331.

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.

Table 1

Student Cohort by Bridges Implementation Phase

	Grade by Academic Year					
Student Cohort	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	Bridges Implementation Phase
	3	4	5			
1	3	4	5			
	3	4	5			
		3	4	5		1 (2016-17)
2		3	4	5		2 (2017-18)
		3	4	5		3 (2018-19)
			3	4	5	1 (2016-17)
3			3	4	5	2 (2017-18)
			3	4	5	3 (2018-19)

Note. Shaded boxes indicate periods in which *Bridges* was implemented.

Table 2

Percentages of Demographic and English Language Proficiency Composition Across Treatment Groups

Variable	*Investigations* (*n* = 3,354)	*Bridges* (*n* = 1,839)	Difference *p*-value
Student with Individualized Education Plan (IEP)	14%	15%	.153
Free or Reduced-Price Lunch Eligible	50%	51%	.300
Female	50%	50%	.765
Race/Ethnicity			
White	42%	42%	.860
Black or African American	17%	18%	.186
Hispanic/Latino	22%	23%	.802
Asian or Asian American	9%	8%	.203
Native Hawaiian/Pacific Islander or American Indian/Alaska Native	0%	0%	.497
Multiracial	10%	9%	.314
Parent Education			
Less than High School Degree	7%	6%	.077
High School Degree	18%	19%	.344
Some College or Technical Degree	23%	24%	.336
Four-Year Degree	17%	19%	.016
Graduate School/Professional Degree	31%	28%	.052
Missing Education Level	5%	4%	.097
English Language Proficiency			
Proficient (Levels 6–7)	77%	78%	.681
Mid English Proficiency (Levels 4–5)	16%	12%	< .001
Low English Proficiency (Levels 1–3)	7%	11%	< .001

Note. Parent education is the highest- or only-reported level across third, fourth, and fifth grade. Some estimates rounded to 0 due to small sample sizes. Two-tailed proportion test used to calculate treatment balance (tests not corrected for clustering). Using a multinomial logistic regression to predict ELP level using treatment condition, *Bridges* students are more likely to have low ELP (compared to mid), but this effect is removed when controlling for cohort year.

Table 3

Descriptive Statistics of Assessment Data Across Treatment Groups

Variable	Mean (Standard Deviation) *Investigations*	Mean (Standard Deviation) *Bridges*	Mean Difference *p*-Value
MAP Math RIT Score Fall	209.64 (17.52)	210.72 (17.56)	.485
MAP Reading RIT Score Fall	206.24 (18.22)	206.72 (17.82)	.744
MAP Math RIT Score Spring	219.22 (19.01)	221.37 (19.06)	.201
MAP Reading RIT Score Spring	212.70 (17.33)	213.22 (16.93)	.709
ACCESS Overall Composite Score Proficiency Level (Grade 4)[a]	4.7 (0.97)/**4.6**	4.2 (0.95)/**4.1**	.004

Note. All data from grade five except ACCESS. 3,345 control and 1,836 *Bridges* students had spring MAP reading data. 440 *Bridges* students and 927 control students had ACCESS scores. Tests of condition differences clustered at school by cohort level (87 clusters for MAP, 86 for ACCESS). Since we use a DiD design, we also report MAP reading and math data from the spring of fifth grade. Treatment and comparison groups score similarly in the spring on both measures. However, these data are meant only to describe the sample characteristics in achievement level and are not tests of our main research questions. The mean difference in ACCESS scores is significantly reduced and not statistically significant when controlling for cohort year. All values are unweighted (see Missing Data section in Methods for more information on weighting for missingness).

[a]Median values presented in bold as the ACCESS measure exhibits a somewhat nonnormal distribution.

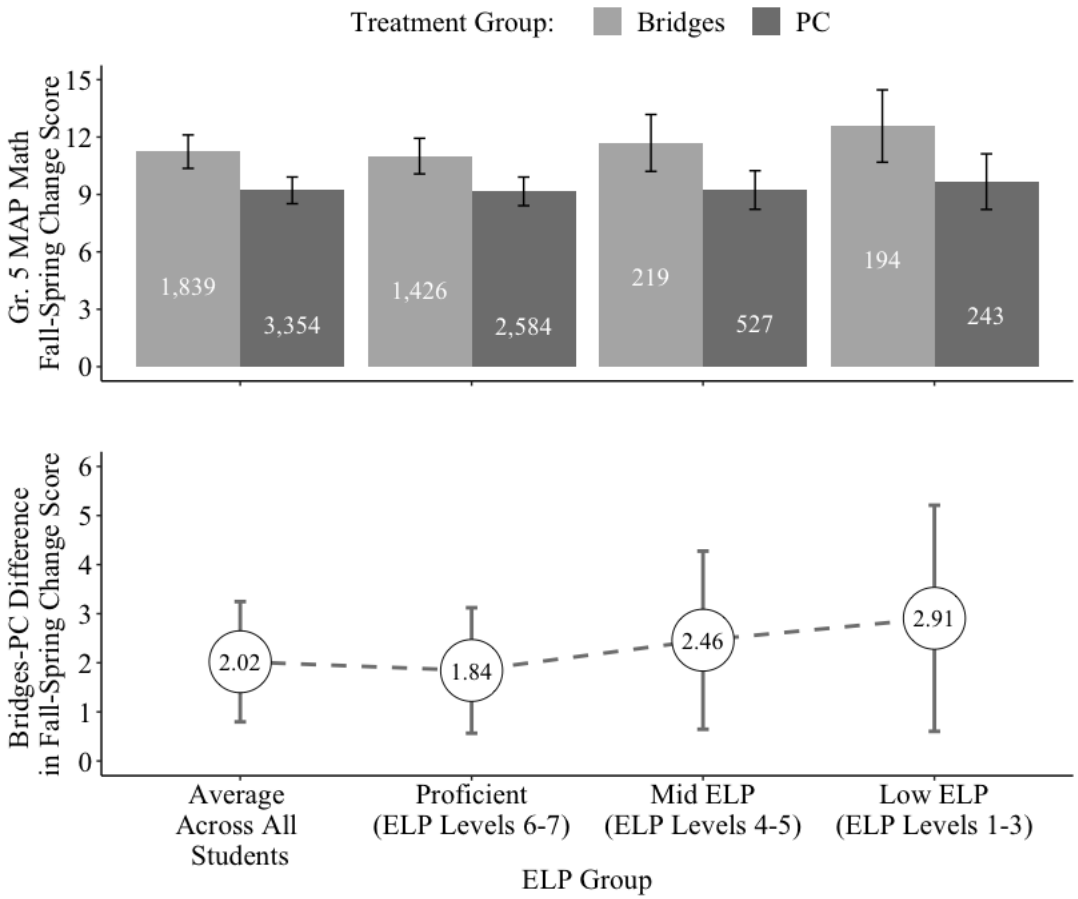
Figure Captions

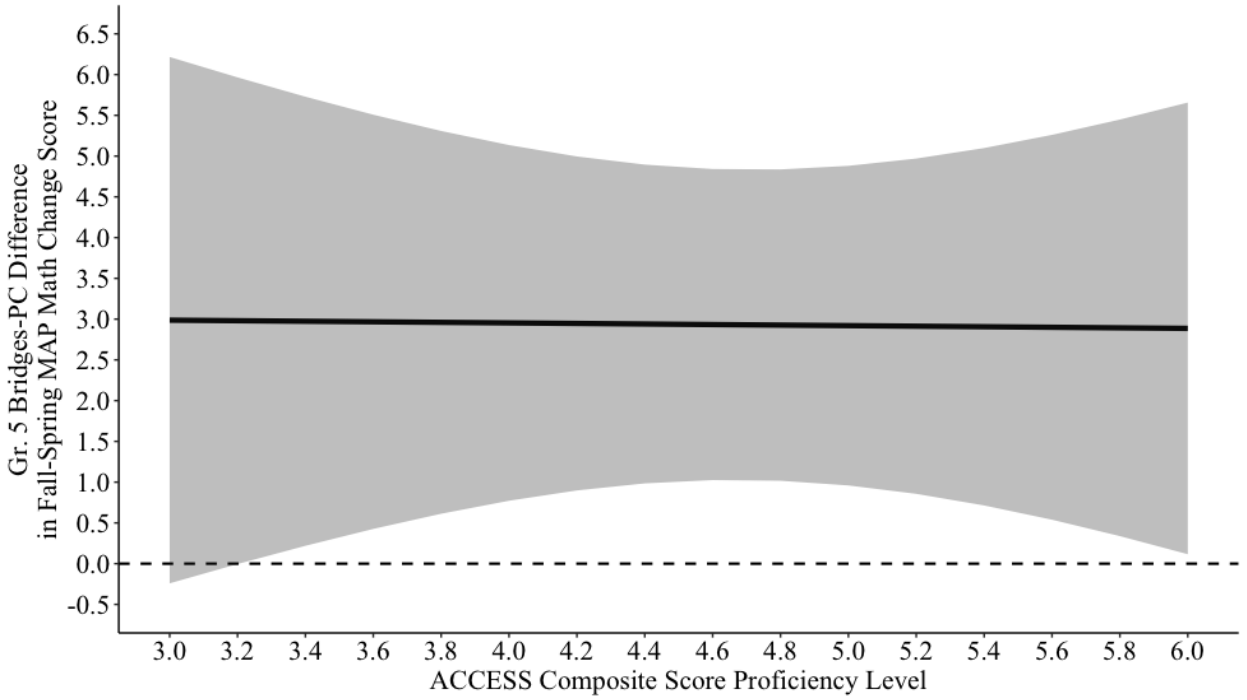
Figure 1. Predictive margins of *Bridges* and PC groups across ELP levels (top) with group sample sizes and DiD simple effects coefficients (bottom).

Note. Error bars represent cluster-robust 95% confidence intervals. Figure produced in *ggplot2* (Wickham, 2016). PC = Previous Curriculum.

Figure 2. *Bridges* impact estimates across ACCESS Overall Composite Score proficiency levels 3 to 6

Note. Error bars represent cluster-robust 95% confidence intervals. Figure produced in *ggplot2* (Wickham, 2016). PC = Previous Curriculum.


Treatment Group:
Bridges
PC
Gr. 5 MAP Math Fall-Spring Change Score
15
12
9
6
3
0
1,839
3,354
1,426
2,584
219
527
194
243
Bridges-PC Difference in Fall-Spring Change Score
6
5
4
3
2
1
0
2.02
1.84
2.46
2.91
Average Across All Students
Proficient (ELP Levels 6-7)
Mid ELP (ELP Levels 4-5)
Low ELP (ELP Levels 1-3)
ELP Group


Gr. 5 Bridges-PC Difference
in Fall-Spring MAP Math Change Score
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0
5.2
5.4
5.6
5.8
6.0
ACCESS Composite Score Proficiency Level

Supplemental Materials for *Examining Bridges in Mathematics and Differential Effects Among English Language Learners*

In this supplemental materials document, we present additional information regarding inverse probability weighting for missing data, descriptive and inferential statistics, regression model tables, as well as robustness checks, a sensitivity analysis, and additional results using grade five ELP data. Table S1 displays additional information on unadjusted and adjusted change score estimates and effect sizes; Table S2 displays the Benjamini-Hochberg Correction table for our primary results; and Tables S3 and S4 display the full regression results from Models 1 and 2, respectively (which are the basis for the marginal effects we present in the main text). Figures S1 and S2 display results of robustness checks discussed in the main text. Finally, Figure S3 displays adjusted estimates and marginal effect results of Models 1 and 2 estimated using ELP levels and ACCESS scores measured in fifth grade.

Model Used for Constructing Inverse Probability Weights

We used a logistic regression model to construct inverse probability weights (IPWs; Seaman & White, 2013) to account for missing data in our analysis of the impact of *Bridges* on annual MAP math growth in fifth grade. This is a single-level model with cluster-robust standard errors; however, we use *i* and *j* to denote student versus school-level variables. The model is as follows:

$$\begin{aligned} logit\ MAP\ Change\ Score\ Observed_{ij} = {} & \beta_0 + \beta_1\ Bridges_j + \beta_2\ ELP\ Level_{ij} + \\ & \beta_3\ Fall\ MAP\ Reading_{ij} + \beta_4 Special\ Ed._{ij} + \beta_5\ FRL_{ij} + \beta_6\ Female_{ij} + \\ & \beta_7\ Parent\ Education\ Level_{ij} + \beta_8\ Cohort\ Year_{ij} + \\ & \beta_9\ Technology\ Implementation\ Cohort_j + \beta_{10}\ Bridges_j \times X_{ij} + \varepsilon_{ij} \end{aligned}$$

where $MAP\ Change\ Score\ Observed_{ij}$ is a binary variable coded 1 if a student has both fall and spring MAP math in grade five and 0 if not. We used five indicators for six-category race/ethnicity (Native Hawaiian/Pacific Islander and American Indian/Alaska Native combined

into one indicator). We included five indicators of parent education level (no high school, some college/technical degree, college degree, graduate/professional degree, missing education level) with high school completion as the reference group. As reported in the main text, we use the highest-reported education level across grades three to five (or the level that is available). Cohort year is a three-category variable representing each of the student data cohorts (i.e., students in grade five in 2015, 2016, or 2017) with cohort 3 (2017) as the reference group. Technology implementation cohort is a six-category variable indicating whether a school (regardless of cohort year) was implementing or planned to implement a one-to-one student technology plan that overlapped with Bridges implementation. Five indicator variables were included and the first cohort of implementation, which occurred in 2015-16, was used as the reference group. β_{10} represents interaction terms between treatment condition and each of the covariates (represented as X_{ij}) except each technology cohort indicator. Bridges is interacted with only the year cohort 2 (2016-17) indicator, as no cohort 1 (2015-16) students received Bridges. ε_{ij} is a random error term clustered at the school by cohort level. We then constructed IPWs by obtaining predicted probabilities from the logit model and dividing 1 by the predicted probability of being a complete case. The weights sum to the number of individuals with complete data on the covariates (n = 5,359)

Supplemental Information on Primary Analysis Results

Supplemental Table S1 displays unadjusted (unweighted) and adjusted (weighted) change score means for each treatment group from Models 1 and 2 along with sample sizes for each group, average marginal effects for each comparison, and the corresponding Hedges's *g* estimate (standardized on change score standard deviations).

Supplemental Table S1.
Unadjusted and adjusted change score estimates by treatment and ELP group

	Unadjusted Unweighted Means (*SD*)	Adjusted Weighted Means	*n*	Estimated Difference (*SE*)	Effect Size in Change Score *SD* (Hedges's *g*)
All Students					
Bridges	10.647 (7.882)	11.236	1,839	2.022	0.250
PC	9.580 (8.197)	9.215	3,354	(0.616)	
English Language Proficient					
Bridges	10.553 (7.843)	11.005	1,426	1.843	0.226
PC	9.660 (8.339)	9.162	2,584	(0.643)	
Mid ELP					
Bridges	10.909 (7.170)	11.693	219	2.461	0.336
PC	9.493 (7.375)	9.232	527	(0.913)	
Low ELP					
Bridges	11.036 (8.896)	12.576	194	2.906	0.337
PC	8.914 (8.372)	9.670	243	(1.159)	
Subgroup with ACCESS Scores[a]					
Bridges	11.172 (7.960)	12.089	440	2.935	0.376
PC	9.605 (7.719)	9.154	927	(0.964)	

Note. Each effect size calculated using the *SD* and *n* from each treatment group within each ELP level. Effect sizes calculated using Stata's *esizei* command.
[a]Adjusted estimates from Model 2.

Supplemental Table S2 below displays the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) that we applied to our primary hypothesis tests to control for the false discovery rate (FDR). We adopt the procedure for implementing the Benjamini-Hochberg correction described in What Works Clearinghouse (2020b). The first column, Term, displays each of the primary tests we considered in our main analysis using Models 1 and 2 (described in the main text). The second column shows the observed *p*-value of the main and differential effects tests (all of which are reported in the main text). Using the conventional alpha level of .05, the alpha critical value is divided by the ranking of the observed *p*-value for each test. The highest observed *p*-value that is less than the BH-adjusted critical value is the new criterion for significance. In this case, both significant terms (the main effects of *Bridges* from Models 1 and 2) remained significant after adjustment.

Supplemental Table S2
Benjamini-Hochberg (BH) Corrections for Primary Statistical Tests

Term	Observed *p*-value of Test	*p*-value Rank	BH Adjusted Critical Value	Significant After Adjustment?
Bridges Main Effect Model 1	.001*	1	.007	Yes
Bridges Main Effect Model 2	.003*	2	.014	Yes
Bridges Differential Effects				
Low/Mid Combined vs. Proficient	.273	3	.021	No
Low ELP vs. Proficient	.324	4	.029	No
Mid ELP vs. Proficient	.464	5	.036	No
Low ELP vs. Mid ELP	.724	6	.043	No
Across ACCESS Score	.964	7	.050	No

Note. The formula for the adjusted critical value is .05* [Rank/7]. *Significant at .05 level prior to adjustment.

Supplemental Tables S3 and S4 display the full regression results of Models 1 and 2, respectively. Adjusted means and average marginal effects (AMEs) presented in the main paper (above in Table S2) are based on these two models. Importantly, the effects in these tables are not the primary estimates of interest. The effects of interest in the main text, the AMEs, represent the effect of *Bridges* (as well as the variation in the *Bridges* effect across ELP) averaged across all covariates and interactions in the model (Williams, 2012). The coefficient for *Bridges* in Table S3 below, on the other hand, would represent the effect of *Bridges* among students with English-proficient ELP levels within only cohort 3 (because there is an interaction between *Bridges* and cohort 2 students). This is not the estimate of interest, because this in fact represents the simple effect of *Bridges* among English-proficient students in cohort 3. The primary effect of interest is the main effect of *Bridges* averaged across all other terms in the model, which is what we report in the main text. Similarly, the simple effects of *Bridges* at each ELP level reported in the main text (produced using the *at*—rather than *over*—command within *margins* in Stata) are averaged across all other terms in the model. The interaction coefficients between *Bridges* and ELP level in Table S3 represent the variation in the treatment effect among only cohort 3 students (due to the interaction between cohort 2 and *Bridges*). This same logic applies to Table S4 (the regression results for Model 2 from the main text). We report these results to demonstrate the overall models that were used to calculate the AMEs we calculated in the main text. For more information on how AMEs in Stata are calculated, see Williams (2012). All standard errors are adjusted for clustering at the school-by-cohort level following on recent recommendations to cluster at the level of treatment assignment (Abadie et al., 2017), which in this case was schools within year. We used Stata's *vce(unconditional)* option within the *margins* command to calculate standard errors for AMEs.

Supplemental Table S3.

Unstandardized Regression Coefficients for Bridges Main Impact Model (Model 1)

Predictor	Est.	Cluster-Robust SE	*t*	*p*	95% Confidence Interval Low	95% Confidence Interval High
Bridges	2.379	0.772	3.080	.003	0.844	3.914
Mid ELP	0.070	0.508	0.140	.890	-0.939	1.080
Low ELP	0.508	0.725	0.700	.485	-0.934	1.950
Bridges X Mid ELP	0.617	0.839	0.740	.464	-1.051	2.285
Bridges X Low ELP	1.063	1.072	0.990	.324	-1.068	3.194
Less Than High School Degree	-0.092	0.476	0.190	.847	-1.038	0.853
Some College/Tech Degree	0.284	0.372	0.760	.447	-0.455	1.024
Four-Year Degree	0.592	0.463	1.280	.204	-0.328	1.513
Graduate/Professional Degree	1.129	0.421	2.680	.009	0.293	1.965
Missing Education Level	-0.428	0.625	0.680	.496	-1.670	0.815
Black or African American	-0.919	0.416	2.210	.030	-1.747	-0.092
Hispanic/Latino	-0.505	0.437	1.150	.252	-1.374	0.365
Asian	0.206	0.437	0.470	.639	-0.663	1.074
Multiracial	-1.339	0.476	2.810	.006	-2.285	-0.392
Native Hawaiian/Pacific Island or American Indian/Alaskan Native	-0.023	2.005	0.010	.991	-4.009	3.962
Fall Gr. 5 MAP Reading	0.015	0.010	1.470	.146	-0.005	0.036
Free/Reduced Price Lunch	-0.179	0.371	0.480	.631	-0.917	0.559
Student with Individualized Education Plan	-0.924	0.584	1.580	.117	-2.086	0.237
Female	-0.210	0.225	0.940	.352	-0.657	0.236
Cohort 1 (15-16)	1.455	0.613	2.370	.020	0.237	2.674
Cohort 2 (16-17)	2.782	0.713	3.900	.000	1.365	4.200
Bridges X Cohort 2	-1.632	0.956	1.710	.092	-3.533	0.269
Technology Plan Cohort 2	1.497	1.013	1.480	.143	-0.517	3.511

Technology Plan Cohort 3	-0.640	0.653	0.980	.330	-1.939	0.659
Technology Plan Cohort 4	-1.287	0.868	1.480	.142	-3.013	0.439
Technology Plan Cohort 5	-0.068	0.780	0.090	.931	-1.619	1.484
Technology Plan Cohort 6	-0.482	0.889	0.540	.589	-2.249	1.286
Intercept	5.348	2.206	2.420	.017	0.963	9.732

Note. n = 5,193. Standard errors (SEs) corrected for 87 school-by-cohort clusters. ELP reference group is proficient. Parent education level reference group is high school.

Supplemental Table S4.
Unstandardized Regression Coefficients for Secondary Bridges Impact Model (Model 2)

Predictor	Est.	Cluster-Robust SE	*t*	*p*	95% Confidence Interval Low	95% Confidence Interval High
Bridges	4.521	3.570	1.270	.209	-2.578	11.619
ELP	-0.114	0.382	0.300	.766	-0.873	0.645
Bridges X ELP	-0.034	0.775	0.040	.965	-1.574	1.507
Less Than High School Degree	-0.353	0.512	0.690	.492	-1.372	0.665
Some College/Tech Degree	0.268	0.521	0.510	.608	-0.768	1.303
Four-Year Degree	-0.077	0.937	0.080	.935	-1.940	1.786
Graduate/Professional Degree	1.216	0.802	1.520	.133	-0.378	2.810
Missing Education Level	-1.001	0.873	1.150	.255	-2.738	0.736
White	2.074	1.131	1.830	.070	-0.174	4.323
Black/African American	0.621	0.958	0.650	.519	-1.284	2.526
Asian	0.661	0.678	0.970	.333	-0.687	2.009
Multiracial, Native Hawaiian/Pacific Island, or American Indian/Alaska Native	2.220	1.239	1.790	.077	-0.244	4.683
Free/Reduced Price Lunch	1.281	0.544	2.350	.021	0.199	2.362
Student with Individualized Education Plan	-2.283	0.912	2.500	.014	-4.096	-0.470
Fall Grade 5 MAP Reading	-0.012	0.023	0.540	.588	-0.058	0.033
Female	-0.222	0.404	0.550	.585	-1.025	0.582
Cohort 1 (15-16)	2.640	1.118	2.360	.020	0.417	4.862
Cohort 2 (16-17)	4.103	1.154	3.560	.001	1.809	6.397
Bridges X Cohort 2	-3.814	1.674	2.280	.025	-7.143	-0.485
Technology Plan Cohort 2	0.465	1.038	0.450	.656	-1.600	2.529
Technology Plan Cohort 3	-0.491	1.128	0.440	.664	-2.733	1.751
Technology Plan Cohort 4	-0.415	0.948	0.440	.663	-2.301	1.470
Technology Plan Cohort 5	0.303	0.899	0.340	.736	-1.483	2.090
Technology Plan Cohort 6	0.467	1.161	0.400	.689	-1.842	2.776

Intercept	8.594	3.942	2.180	.032	0.757	16.431

Note. n =1,367. Standard errors (SEs) corrected for 86 school-by-cohort clusters (the main model corrects for 87). Multiracial students and Native Hawaiian/Pacific Islander or American Indian/Alaska Native students collapsed into a single indicator variable due to small sample sizes. Race/ethnicity reference group is Hispanic/Latino. ELP = ACCESS composite proficiency score.

Robustness Checks

As we described in the main text, we conducted a number of additional analyses to assess the robustness of our models and the sensitivity of our results to unmeasured confounders. We tested the robustness of Model 1 to alternative specification, assessed the assumptions of DiD and robustness of our main finding using an analysis of parallel trends and falsification tests (St. Clair & Cook, 2015; Furquim et al., 2020), conducted a sensitivity analysis of the main effect of *Bridges*, and assessed our findings using grade five ELP/ACCESS data. Below, we provide a description of those analyses

Assumption Checks and Alternative Model Specifications

We addressed the assumptions of the regression model by inspecting residuals and potential outlying or influential data points. Robust clustered standard errors in all analyses account for the clustered structure of the data (students nested in schools) and heteroskedasticity. Removal of large and outlying positive or negative change scores slightly reduced the reported main effect point estimates in Models 1 and 2, though simple effects of both models (and by extension the moderation estimates) were more sensitive to these outliers (the low ELP simple effect decreased by approximately 0.40 and the Bridges moderation estimate between ACCESS scores of 3 and 6 increased by approximately 1.70). However, moderation estimates remained imprecise and nonsignificant, and all main effects remained similarly significant. As a result, we retained outlying data points in the analyses given the limited impact on our inferences and to represent the full distribution of change scores in the analyses.

We inspected the robustness of the primary results to different model and error specifications. Errors clustered at the school-level regardless of cohort (i.e., 29 schools) were only slightly smaller than the reported models. Point estimates of Model 1 varied across

multilevel models with random intercepts at the school-level (b = 1.678, SE = .568, z = 2.95, p = .003) or school by cohort level (b = 2.036, SE = .705, z = 2.89, p = .004). A single-level analysis of covariance (ANCOVA) version of Model 1 with spring MAP math as the outcome, controlling for fall MAP math (with school by cohort clustered errors), showed similar significant effects as well (b = 2.118, SE = 0.593, t = 3.57, p = .001). All alternative models were estimated with the same covariates and IPWs as described for the primary analysis. Impact estimates similarly varied across ELP levels in these alternative models. Including an indicator variable of whether students switched schools between the prior and current year or controlling for prior year special education status (since status might have changed during the year of Bridges for some students) had little effect on the reported main effect estimates. Restricting the sample to those who did not switch schools between prior and current years also had a minimal impact on estimates.

Among only students in Bridges schools (n = 1,839), we tested whether those assigned to Bridges schools in grade four and continued in a Bridges school in grade five (n = 558) showed different treatment effects than students who received Bridges only in grade five (n = 1,281) (controlling for identical covariates except enrollment year indicator for 2015 and using the same IPWs). Students who received two years of Bridges instruction did not change differently in fifth grade than those who received one year (b = -0.281, SE = 0.873, t = 0.32, p = .750). This effect did not vary significantly by ELP level, though a pattern of point estimates similar to Model 1 emerged (low ELP b = 1.385, mid ELP b = 0.598, proficient b = -0.648). These results suggest our primary findings may not be only a result of length of exposure to *Bridges.*

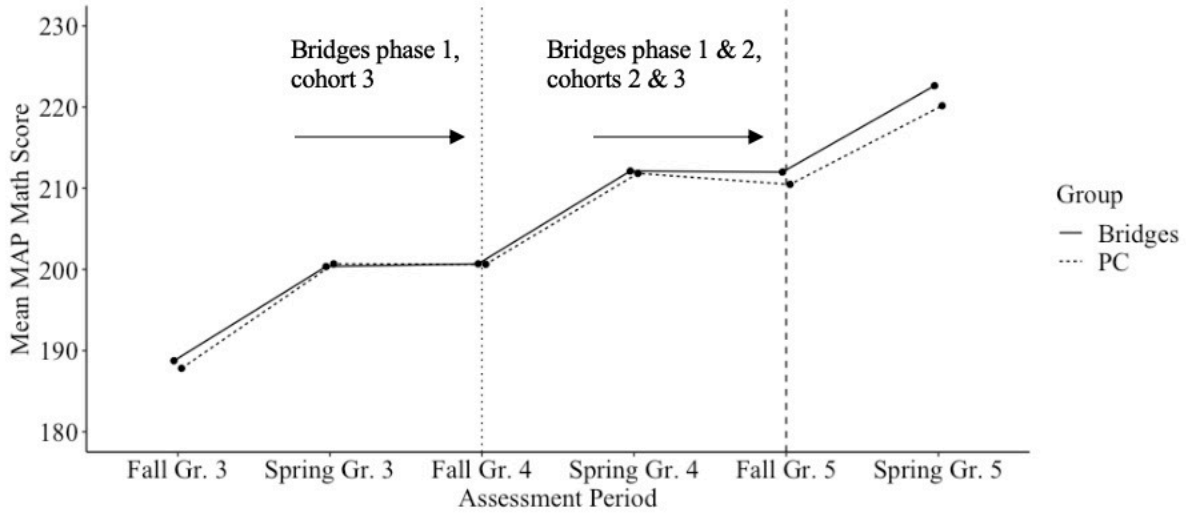
Parallel Trends Analysis and Falsification Tests

Figure S1 displays MAP math mean scores across assessment periods between grades

three and five. This graph is a representation of the parallel trends assumption of DiD (St. Clair & Cook, 2015). Violation of this assumption may impact the internal validity of the design if there was systematic variation between treatment and comparison conditions prior to the treatment period (Lechner, 2011; St. Clair & Cook, 2015). This graph suggests minimal variation between treatment and comparison groups prior to treatment implementation, particularly prior to implementation in grade five. Only students who had complete MAP math data across all waves are included (regardless of if they had MAP reading data).

Figure S1

Mean MAP math achievement estimates in each assessment period in grades three to five.



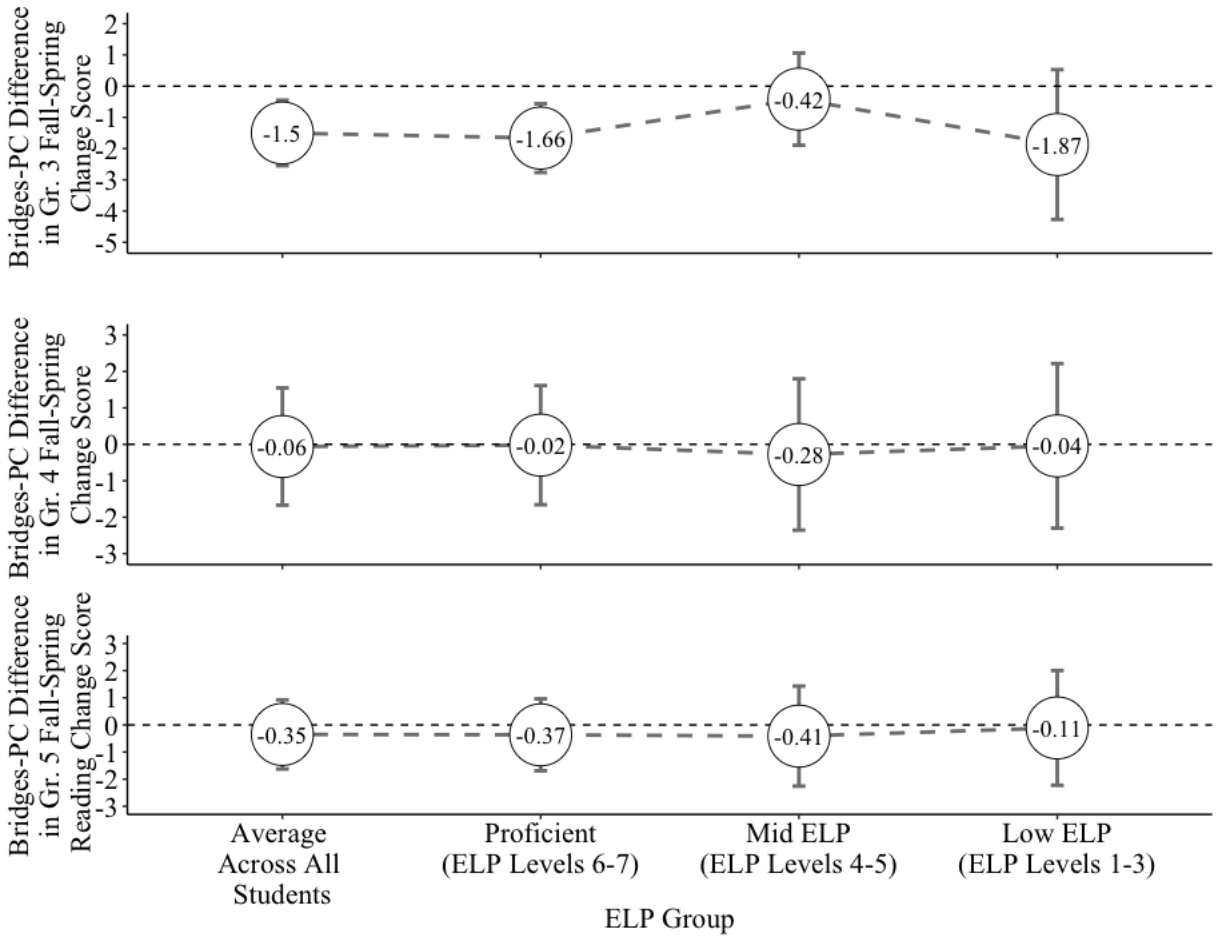
Note. Means calculated for students with complete MAP math data across all assessment periods (total n = 4,589, *Bridges* n = 1,637). Adjusted estimates from multilevel models controlling for grade five enrollment year show similar results. Plot produced in *ggplot2* (Wickham, 2016). Points slightly offset from each other for visual clarity.

Figure S2 below displays falsification or nonequivalent outcomes tests (St. Clair & Cook, 2015; Furquim et al., 2020) of *Bridges* across all students and within each ELP level. We use models identical to Model—including the same clustered standard errors, covariates, and IPWs—but we replace our main outcome (grade five MAP mathematics change score) with the relevant falsification or nonequivalent outcome. When our outcome is MAP reading change score, we remove fall grade five MAP reading as a covariate (and including MAP fall grade five mathematics minimally changes estimates). In the case of measuring the treatment on time periods prior to the treatment (the top two panels of Figure S2), these tests help further assess whether treatment and comparison differences existed prior to the main treatment period (conditioning on the covariates we used in the primary analysis using Model 1), which may threaten the internal validity of our design. The nonequivalent outcomes test (bottom panel of Figure S2) helps establish whether treatment-control differences were present on outcomes that may not be reasonably expected to be impacted by the treatment to the same extent as the primary outcome. Although it is possible that a mathematics intervention could impact reading skills (and that may be of substantive interest), our primary outcome of interest and hypothesized effects were focused on mathematics. Only students who were observed in the primary analysis model (Model 1) and had data on the relevant outcome measure, covariates, and had IPWs are included in these analyses. Grade four MAP change scores did not differ in statistical significance or magnitude between treatment and comparison groups, nor did grade five MAP reading change scores. Grade three MAP change scores, however, were statistically significantly lower on average between treatment and comparison groups. This trend is somewhat apparent on Figure S1 as well. This may suggest a departure from parallel trends that could affect the internal

validity of the design. However, this trend was apparent two years prior to the majority of schools implementing *Bridges* in our design (though some schools implemented in grade four).

Figure S2

Falsification tests of main treatment effects and simple effects across ELP levels.



Note. Grade three MAP math change score placebo test (top panel, total n = 4,657), grade four MAP math change score placebo test (middle panel, total n = 4,984), and grade five MAP reading change nonequivalent outcomes test (bottom panel, total n = 5,181). These models include the same covariates as Model 1 except the model with MAP reading as the outcome (bottom panel) does not include fall grade five MAP reading as a covariate. Error bars represent

cluster-robust 95% confidence intervals (using fifth grade school-by-cohort clusters). PC = previous curriculum. Plot produced in *ggplot2* (Wickham, 2016).

Sensitivity Analysis

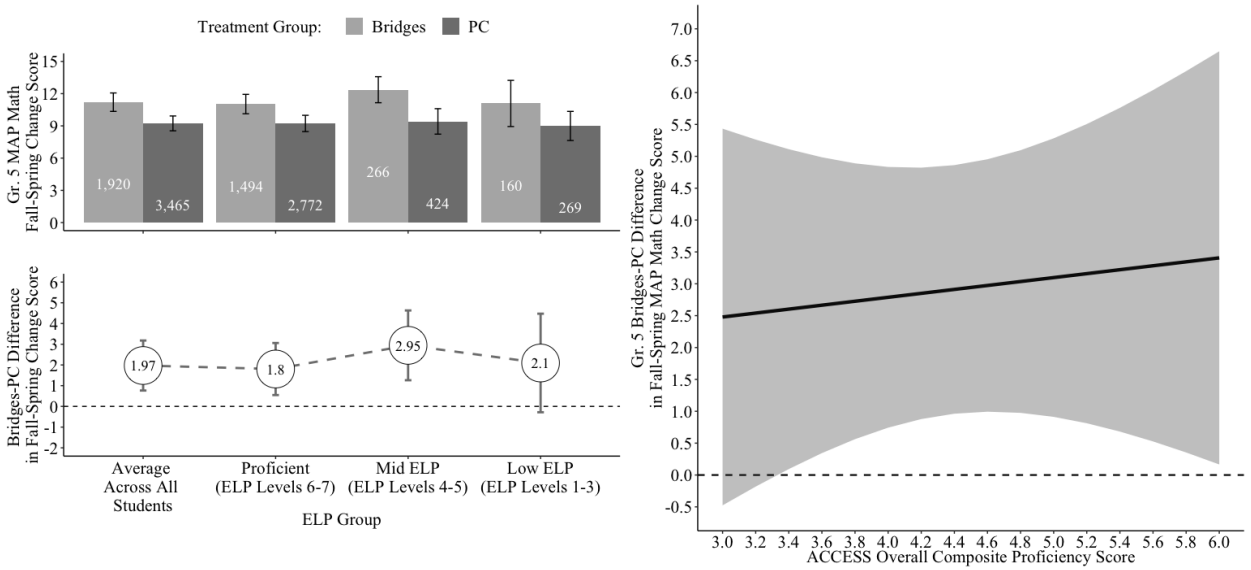
We conducted a sensitivity analysis of the main effect of *Bridges* (b = 2.022, SE = 0.616) to determine the extent to which an unobserved variable would need to correlate with our outcome and treatment assignment to produce a nonsignificant effect of *Bridges*. We used the R Shiny App *Konfound-It!* (Rosenberg et al., 2018) for this procedure. We used the main effect reported in the main text (b = 2.022), its cluster-robust standard error (0.616), 5,193 observations, and the number of covariate terms in Model 1 (26) to complete this procedure. The results of this procedure indicate that an unmeasured variable would need correlate greater than .137 with both *Bridges* assignment as well as grade five MAP mathematics change scores while accounting for covariates (and thus have an impact of .019 [i.e., .137*.137]) to render our observed effect nonsignificant (Frank, 2000).

Analyses using Grade Five ELP and ACCESS

Figure S3 displays adjusted estimates and marginal effects of Models 1 and 2 estimated using grade five ELP or ACCESS data (rather than prior year). We draw the same conclusions from these estimates as we do from the primary models in the main text. The main effect of *Bridges* in the left panel is statistically significant and of similar magnitude (g = 0.243) to the main model presented in the main text. This effect does not vary significantly across ELP levels (left panel), nor do *Bridges* estimates vary significantly ACCESS scores (right panel).

Figure S3

Bridges impact estimates using ELP measured in grade five.



Note. Left graph, top panel displays covariate-adjusted mean change scores for each group. Left graph, bottom panel displays the estimated difference between each mean in the top panel (the AMEs of *Bridges*). The graph on the right displays the AME of *Bridges* at representative values of ACCESS scores. Sample sizes differ from main analyses since more student data were available using grade five ELP (full n = 5,385), though there were fewer students with ACCESS scores (ACCESS n = 1,161). Models used to estimate these effects are identical to Models 1 and 2 reported in the main text (except for the different ELP or ACCESS variable). We constructed IPWs for these analyses using a model identical to what was used for the primary analyses (see first page of this document) with the exception of using grade five ELP level. ACCESS scores and ELP data were recorded in the winter during fifth grade (i.e., after all *Bridges* schools started implementation) and thus may constitute a posttreatment variable for all *Bridges* students (Montgomery et al., 2018). This would only be problematic in the main analyses Plot produced in *ggplot2* (Wickham, 2016).

References

Abadie, A., Athey, S., Imbens, G. W., & Woolridge, J. (2017). *When should you adjust standard errors for clustering?* National Bureau of Economic Research Working Paper No. 24003. https://www.nber.org/papers/w24003

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300.

Furquim, F. F., Corral, D., & Hillman, N. W. (2020). *A primer for interpreting and designing difference-in-differences studies in higher education research*. In L. W. Perna (Ed.), Higher Education: Handbook of Theory and Research Vol.:XXXV. Springer.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165-224. doi: 10.1561/0800000014

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on post-treatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775. doi: 10.1111/ajps.12357

Rosenberg, J. M., Xu, R., & Frank, K. A. (2018). *Konfound-It!: Quantify the robustness of causal inferences*. http://konfound-it.com

St. Clair, T., & Cook, T. D. (2015). Difference-in-difference methods in public finance. *National Tax Journal, 68*(2)*,* 319-338. doi: 10.17310/ntj.2015.2.04

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. 2016.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, *12*(2), 308-331.

What Works Clearinghouse. (2020b). What Works Clearinghouse procedures handbook, version 4.1. Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf