A Year of Desirable Difficulties:* The Impact of Interleaving Math Practice in Nigeria

Michael Kremer† Guthrie Gray-Lobe‡ Joost de Laat§
Lotte van der Haar¶

September 29, 2026

This paper studies whether reorganizing classroom practice improves learning. In a year-long school-randomized experiment in Nigerian primary schools, students received blocked mathematics practice, grouping problems that use the same strategy, or interleaved practice, mixing current and previously taught problem types. Interleaving increased termly assessment scores by 0.28 standard deviations, with larger gains for initially lower-achieving students, but produced no detectable effect on a cumulative end-of-year assessment. A simple model of retrieval difficulty offers one possible interpretation: early gains may make later practice easier and less productive. The results highlight the promise and limits of changes to practice design.

JEL Classification: I20, I21, I25, C93.

*This study reports results from an experiment designed and conducted in schools operated by New-Globe. We thank Shannon May, Sean Geraghty, Tim Sullivan, Daniel Rodriguez-Segura, and Clotilde de Maricourt for their collaboration and for providing valuable feedback. We also thank Jim Heckman, Paul von Hippel, Tobias Richter, Doug Rohrer, Elakiya Ananthakrishnan and the participants of the 2024 Midwest International Economic Development Conference for valuable feedback. The evaluation received support from the J-PAL Post-Primary Education Initiative and the International Growth Centre (IGC). This study was determined to be Not Human Subjects Research under federal regulations [45 CFR 46.102(f); 21 CFR 50.3(g)] by the University of Chicago Social and Behavioral Sciences Institutional Review Board (IRB20-2191) and by the Harvard University Institutional Review Board (IRB18-2108). The order of author names was randomized using the American Economic Association Author Randomization Tool (Link).

†University of Chicago, email: kremermr@uchicago.edu
‡University of Chicago, email: graylobe@uchicago.edu
§Wageningen University, email: joost.delaat@wur.nl
¶Wageningen University, email: lotte.vanderhaar@wur.nl

1 Introduction

How students practice is often determined by defaults embedded in textbooks, teaching guides, and other instructional materials. Cognitive science suggests that these defaults may be improved by incorporating "desirable difficulties," such as spacing practice over time, interleaving different types of problems, varying the context in which material is encountered, and requiring students to retrieve prior knowledge rather than simply restudying it (Bjork, 1994; Roediger and Pyc, 2012; Dunlosky et al., 2013). These strategies make practice feel slower or more effortful in the short run but can improve subsequent retention and transfer. Students and teachers often avoid them precisely because easier forms of practice generate a stronger immediate impression of progress. Changing the default order of practice may therefore provide a low-cost and politically uncontroversial way to help students study more effectively without increasing instructional time or changing curricular content (Brown et al., 2014; Rohrer et al., 2020a).

Most evidence supporting desirable difficulties comes from relatively short, tightly controlled studies in which researchers manage students' exposure to different forms of practice. Studies of interleaved mathematics practice generally find substantial improvements in retention and transfer. Under blocked practice, students solve several problems of the same type consecutively, so the surrounding problems help signal which solution strategy to use. Under interleaved practice, different problem types are mixed, requiring students to identify the appropriate strategy for each problem. However, these interventions have typically lasted only days or weeks and have been conducted largely in high-income settings (Brunmair and Richter, 2019). Much less is known about their cumulative effects when they are incorporated into routine instruction over an academic year. Over longer periods, students repeatedly revisit previously taught material through later lessons, revision, homework, and assessments. These subsequent learning opportunities may affect whether an initial advantage from more productive practice persists.

We study a year-long, school-randomized experiment in grade-five classrooms in Nigeria. Schools were assigned to blocked or interleaved mathematics practice. In blocked classrooms, students practiced problems requiring the same solution strategy as the lesson taught that day. In interleaved classrooms, the same number of practice problems mixed material from the current lesson with previously taught problem types, requiring students to retrieve prior knowledge and identify the appropriate solution strategy. The intervention therefore changed the organization of practice while holding instructional time, curricular content, and the number of practice problems fixed. Practitioners designed the program in response to students' difficulties classifying new problems and matching them to appropriate solu-

tion strategies. The design drew on evidence that interleaving can strengthen these skills (Mayfield and Chase, 2002; Taylor and Rohrer, 2010a; Rickard et al., 2008; Rohrer et al., 2014). The program was implemented by the school operator through its ordinary system of centrally prepared teacher guides, without direct researcher implementation or supervision.

Interleaved practice increased performance on termly mathematics assessments by 0.28 standard deviations. The gains were largest among initially lower-achieving students, indicating that a more cognitively demanding organization of practice need not disproportionately disadvantage students who begin with weaker skills. This result concerns the organization of practice, rather than the level of curricular content: both groups studied the same material, but interleaved students received fewer immediate cues about which knowledge or solution strategy to apply. Despite the substantial termly gains, however, interleaving produced no detectable improvement on a cumulative assessment at the end of the program. The estimated cumulative effect was 0.03 standard deviations, and the difference between the termly and cumulative effects was statistically significant.

This empirical pattern resembles the diminishing treatment effects documented for many educational interventions (Bailey et al., 2017, 2020). Our experiment was not designed to distinguish among competing explanations for the smaller cumulative effect. It cannot determine whether the pattern reflects forgetting by students assigned to interleaved practice, catch-up by students assigned to blocked practice, differences in the skills measured across assessments, or behavioral responses by students and teachers. Instead, the findings motivate a model showing how the mechanism underlying desirable difficulties itself can generate a force toward catch-up.

The model considers repeated opportunities to retrieve and apply a particular skill. The difficulty of a retrieval opportunity depends both on the design of practice and on the learner's existing knowledge: the same mixture of problem types may be highly demanding for a novice but nearly effortless for a student who has already mastered the relevant material. Expected learning is single-peaked in retrieval difficulty. Practice that is too easy generates little additional learning, while practice that is excessively difficult is unlikely to result in successful retrieval. Interleaving can increase initial learning by moving practice toward the productive region between these extremes.

The same mechanism also affects the productivity of later practice. When more demanding initial practice generates additional learning, students enter subsequent retrieval opportunities with greater knowledge. Those opportunities are therefore easier and, over the relevant region of the learning function, less productive. Students who learned less initially continue to experience greater retrieval difficulty and can consequently learn more from later opportunities to revisit the material. With sufficiently frequent subsequent practice, the

initial learning advantage can therefore narrow. The model isolates this mechanism for a single skill; over an academic year, analogous processes may operate repeatedly and overlap across the many skills covered by the curriculum.

We connect to a broad literature on the persistence and fadeout of educational interventions, which offers several complementary explanations for why early gains may diminish over time. This literature highlights mechanisms such as forgetting among treated students, limited transfer of acquired skills, differences in the types of skills targeted, measurement issues, and the role of subsequent investments and environments (Carneiro and Heckman, 2003; Cunha and Heckman, 2007; Cunha et al., 2010; Bailey et al., 2017, 2020). Our model provides a formal representation of the idea that control students catch up to treated students, and that this convergence is shaped by the environments both groups enter after the intervention. In particular, it captures how subsequent instructional contexts can differentially affect students depending on their initial skill levels, generating catch-up even when all students face similar post-intervention environments.

In our framework, the productivity of later practice depends on students' prior knowledge, which is itself shaped by the initial treatment. Even when all students face the same high-quality subsequent environment, those who learned less initially may benefit more from later opportunities because these remain more cognitively demanding and therefore more productive. This contrasts with existing accounts that treat the quality of subsequent environments as exogenous. One exception is Claessens et al. (2014), which descriptively examines whether preschool is more effective when treated students receive more advanced instruction in later grades. Our approach differs by providing a mechanism through which initial gains can influence the productivity of later learning opportunities, thereby offering a new lens on the sources of fadeout and suggesting directions for future research.

Simple extensions of the model suggest two potentially important impacts of improvements in the organization of practice toward greater "desirable difficulties." First, productive early difficulty raises knowledge, making later retrieval easier and potentially reducing the practice required to reach mastery. It can also lower the cumulative burden of learning—through fewer failed retrieval attempts or less total effort—when future savings exceed the added initial burden. These margins are rarely observed in evaluations, which typically compare achievement at fixed points in time, but they may matter for motivation, persistence, time allocation, and engagement with future material.

Second, the model also highlights the role of complementary adjustments in instructional pace. When the set of skills taught in a class is held fixed, students who learned less initially due to an inferior organization of practice have opportunities to catch up. To translate accelerated learning into persistent gains in measured achievement, schools may need to

move more rapidly to additional or more advanced content.

The study contributes to a large literature on desirable difficulties and, more specifically, interleaved practice, as well as an emerging literature applying insights from cognitive psychology to the design of scalable, low-cost educational interventions in low- and middle-income countries (Dillon et al., 2017). To our knowledge, this is the first large, multisite field evaluation of an intervention designed specifically to increase students' exposure to desirable difficulties and the first experimental evaluation of interleaving implemented over a full academic year. Prior experimental evidence on mathematics interleaving comes predominantly from researcher-led studies in the United States and Western Europe (see Appendix Table A1). These findings may not generalize to settings with different student populations, baseline learning levels, or complementary educational inputs, and effects achieved in researcher-led pilots may not survive implementation by practitioners at scale (Al-Ubaydli et al., 2019; List, 2022). Our study differs from most prior work not only in duration and scale, but also in implementation: practitioners incorporated interleaving into ordinary classroom instruction through the teaching guides used throughout the school year. The assessments capture retention across a wide range of intervals, from material introduced more than 280 days before testing to material taught only 12 days earlier; on the cumulative assessment, the average interval was approximately 145 days. By comparison, the longest previous experimental evaluation of mathematics interleaving lasted four months and assessed retention one month later (Rohrer et al., 2020b).

The structured-pedagogy setting is important for interpreting implementation. Because practice problems were embedded in centrally prepared teacher guides, the school operator could change the organization of practice without requiring teachers to design new materials themselves. This feature may limit external validity to classrooms in which teachers independently select or write practice problems. At the same time, structured pedagogy is itself widely recommended in international education policy as a cost-effective approach to improving learning in low- and middle-income countries (Akyeampong et al., 2023). Moreover, the relevant implementation margin is not limited to highly scripted school systems. Rohrer et al. (2020a) show that blocked practice remains common in mathematics textbooks, suggesting that interleaving could also be incorporated into textbooks, workbooks, teacher guides, or other curriculum materials used in less structured settings.

The estimated termly effect of 0.28 standard deviations is close to the 0.34 standard deviation effect reported in a meta-analysis of interleaving studies (Brunmair and Richter, 2019). This similarity is informative because standardized effect sizes from short, controlled cognitive-science studies are not necessarily directly comparable with those from education field evaluations. Standard deviations depend on the assessment and sample, and assessments

used in controlled studies may be more tightly aligned with the intervention than the broader tests commonly used in education evaluations (Evans and Yuan, 2022; Angrist et al., 2020). Our results therefore provide evidence that practitioners can operationalize interleaving at scale and reproduce the test-score gains found in more controlled settings. At the same time, the cumulative results and theoretical model suggest that increasing exposure to interleaving alone may not generate persistent achievement gains unless faster learning is accompanied by faster instructional progress.

Beyond its contribution to the interleaving literature, the intervention illustrates a potentially low-cost and politically feasible margin for improving learning. Many education reforms require additional teachers, instructional time, compensation, or infrastructure. Other successful approaches, including tracking and remedial instruction, alter the level or allocation of instruction and can face practical or political constraints. By contrast, desirable difficulties can sometimes be incorporated through default choices in teaching guides, textbooks, and other materials (Rohrer et al., 2020a). They should not be applied universally: the model predicts that excessively difficult practice can harm learning. Nevertheless, even assuming a cost of $13 per child, the estimated termly gains correspond to approximately 2.7 Learning-Adjusted Years of Schooling per $100, comparable to interventions classified by the World Bank as education "Best Buys."

The remainder of this paper is organized as follows: Section 2 describes important features of the setting and program that are relevant for the interpretation of the results; Section 3 describes the methods and empirical framework we use to test the effect of the interleaved practice strategy on math performance; Section 4 presents the results; Section 5 describes a formal model of education production where learning is produced through the difficulty of practice; and Section 6 interprets the results and concludes.

2 Background

This section describes the context and design of the blocked and interleaved practice programs. We begin with the Bridge school setting, mathematics curriculum, classroom practice routines, and assessment schedule, then introduce the desirable-difficulties framework motivating the intervention. We then describe the experimental design, study sample, and test scores used to evaluate the program.

2.1 Setting: Bridge Nigeria

The interleaved versus blocked practice program was implemented in grade five mathematics classrooms across 62 schools in Lagos and Osun states. All participating schools were operated by Bridge Nigeria, a subsidiary of NewGlobe, which manages for-profit private schools primarily in urban informal settlements (henceforth, "Bridge"). NewGlobe also operates private schools in India, Kenya, and Uganda and supports free public schools as a technical partner to governments in Nigeria and Rwanda. Bridge schools typically serve lower-income households. At the time of the study, tuition was approximately eight dollars per student per month.

The interleaving experiment was initiated by Bridge to address student difficulties in classifying mathematics problems, that is, in matching problems to appropriate solution concepts, on cumulative assessments. Bridge staff responsible for designing lesson materials consulted cognitive science research (Taylor and Rohrer, 2010a) to identify methods for improving retention, particularly in categorizing mathematics problems and applying correct solution strategies.

Student exposure to interleaved practice was varied by modifying the centrally standardized materials used to script classroom instruction. Bridge schools follow a structured pedagogy model, where teachers use tablet-based digital guides that script nearly all classroom activities, providing step-by-step instructions for teaching the material. The use of structured pedagogy allows for tighter control over teacher behavior than is typical in classrooms, making this a particularly attractive setting for evaluations of the impact of pedagogical variations.

Bridge students receive two periods of mathematics instruction daily. New material is introduced daily; both periods typically cover similar content and may involve different types of problems.[1] Generally, the first period is more focused on demonstration, while the second period focuses on practice, including some cumulative review. During both 45-minute periods, teachers lead the class in a module known as "my turn, your turn," wherein teachers complete worked examples on the board ("my turn") and then ask students to complete problems in the assignment book ("your turn"). Teachers write all problems on the board, following the electronic teacher guide. Each session includes 16–24 problems focused on a single "topic of the day," with an appropriate solution strategy. The following day introduces a new topic with related practice problems. Each mathematics period includes two "my-turn/your-turn" sessions, covering the same topic with different problems.

Bridge schools administer frequent mathematics assessments throughout the academic

[1]For instance, students might practice a procedure in Mathematics 1 and its inverse in Mathematics 2.

year. Students take up to seven subject-specific tests per year: an initial diagnostic assessment, plus midterm and endterm assessments in each of the three academic terms.[2] These assessments include material from the current term and, in some cases, material taught in previous terms. In Term 3, students also completed an endterm assessment after the cumulative assessment.[3]

2.2 Desirable Difficulties and Interleaved Practice

The intervention builds on a large cognitive science literature showing that some forms of practice improve learning by making practice more effortful in the short run. Bjork (1994) refers to these as "desirable difficulties": learning conditions that increase the difficulty of practice but improve long-run retention or transfer. Three closely related examples are retrieval practice, spacing, and interleaving. Retrieval practice requires learners to recall previously learned material rather than simply restudy it. Spacing distributes repeated exposure to the same material over time rather than concentrating it in a single block. Interleaving mixes different problem types or skills within a practice session, rather than grouping practice by type. In mathematics, blocked practice might present problems in the order aaa-bbb-ccc, while interleaved practice might present them as abc-abc-abc.

Interleaving may be particularly valuable when learning requires students to choose among competing solution strategies. In mathematics, solving a problem requires both executing a procedure and identifying which procedure is appropriate. Blocked practice can make that choice obvious by grouping similar problems together, while interleaving reduces these cues and requires students to retrieve, compare, and classify solution methods. This added short-run difficulty is why interleaving is often described as a desirable difficulty: students may find it harder than blocked practice initially, but it can improve later performance. Consistent with this view, prior studies find that interleaved practice can improve long-run retention and support skills requiring classification, including applying the correct mathematical procedure to a new problem (e.g., Taylor and Rohrer, 2010b; Taylor, 2008; Ziegler and Stern, 2014; Rohrer et al., 2015; Ostrow et al., 2015; Rohrer et al., 2020b).

Interleaving is best understood as a bundle of related learning mechanisms rather than a single treatment contrast. Mixing problem types mechanically creates spacing, because practice on the same type is distributed across time rather than massed together. It also

[2]These assessments provide data for centrally monitoring performance on lesson material.

[3]We do not report impacts on the midterm and endterm assessments because these exams contain items that were taught before the intervention and therefore do not cleanly measure experimentally induced variation in practice opportunities. Instead, we focus on assessments specifically designed to measure skills for which there is experimental variation in practice. We find some evidence of gains on the midterm and endterm assessments, particularly in Term 2, though these effects are not consistently observed in Term 3.

creates retrieval practice, because students must repeatedly retrieve both solution procedures and the mapping between problem features and appropriate strategies. An open question is whether mixing problem types generates benefits beyond the spacing and retrieval practice it creates, for example by improving discrimination across problem categories through a contextual-interference mechanism. We therefore interpret interleaving as a classroom practice that increases exposure to several desirable difficulties at once, rather than as a single mechanism.

The program evaluated in this paper was designed to increase exposure to these difficulties in ordinary classroom settings. Grade five classrooms in Nigerian schools were assigned to receive either blocked mathematics practice, in which practice items tested the same skill in sequence, or a more effortful form of practice that primarily increased interleaving of problem types. Practitioners designed the program based on evidence that interleaved mathematics practice can help students classify new problems and match them to appropriate solution strategies (Mayfield and Chase, 2002; Taylor and Rohrer, 2010a; Rickard et al., 2008; Rohrer et al., 2014). The program was specifically informed by Rohrer (2017), an evidence-based practitioner guide for incorporating interleaving into classroom instruction.

The same framework also matters for interpreting repeated assessment. Tests require students to retrieve previously learned material, and the Bridge assessments described above often draw on material taught at different points in the year. They therefore create distributed retrieval practice and, when they mix items across skills or topics, additional interleaved practice. As a result, end-of-term and end-of-year assessments are not only outcome measures; they may also give students in both experimental conditions additional exposure to difficult, distributed, and mixed retrieval opportunities (Karpicke and Roediger, 2008).

3 Methods

3.1 Procedure

While we refer to the intervention as "interleaving", it simultaneously altered multiple dimensions of practice—including spacing and retrieval effort—and should be interpreted as a bundled increase in exposure to desirable difficulties.

The interleaving program modified the sequence of problems presented to children during the daily "my turn, your turn" module in both Mathematics periods.[4] The number and type of practice problems completed by the teacher during "my turn" were identical in both

[4]In Term 1, only Mathematics 1 practice was blocked in both groups. In Terms 2 and 3, interleaved classrooms received interleaved practice in both Mathematics 1 and 2 sessions.

interleaved and blocked classrooms. The number of practice problems completed by students during "your turn" was also the same in blocked and interleaved classrooms, but the type differed. In blocked classrooms, "your turn" problems were grouped so that all items used the same solution strategy as the "my turn" segment. In interleaved classrooms, "your turn" problems included a mix of aligned and non-aligned items. Aligned problems matched the topic of the day, as in the blocked condition, while non-aligned problems came from other topics covered in the term, often within the same unit (typically spanning 3–5 lessons). Most interleaved practice sets featured roughly equal proportions of aligned and non-aligned problems, though the exact mix varied by lesson.[5] Figure 1 provides examples of problems used in blocked and interleaved classrooms.

In the interleaved condition, mixed-in problems were drawn from lessons that had been taught recently—typically within the preceding three daily lessons or from recent instructional units comprising three to five lessons. Because interleaving was restricted to material already covered within the same term, the first week of each term was identical across treatment and control: both groups practiced only the newly taught topic. As the term progressed, interleaved assignments gradually incorporated a broader range of material, including problems from earlier in the term. This pattern mirrors a common feature of educational testing, in which later assessments generally draw on a wider range of previously taught material and therefore require retrieval over a broader domain of knowledge.

The interleaving program was not designed to isolate the effect of interleaving *per se*. Interleaving manipulates both the spacing (the amount of time between practice attempts) and the degree to which students must discriminate between different categories of problems (Kornell and Bjork, 2008; Taylor and Rohrer, 2010a). Laboratory studies have attempted to disentangle interleaving from related features such as spacing or retrieval practice and found some evidence that there is an effect of interleaving separate from spacing (Taylor and Rohrer, 2010a), but most literature on interleaving views it as a bundle of interventions that manipulate both spacing and the mixture of problem types (Kornell and Bjork, 2008). As such, we view the program as broadly increasing the effort required for practice – consistent with the "desirable difficulties" literature, which encompasses interleaving, retrieval practice, and spaced repetition.

Schools with the interleaving program in grade 5 classrooms also received an unrelated reading program in grade 3 classrooms, adjusting the difficulty of passage reading practice for advanced readers (Aitken et al., 2025).[6] Whether the third-grade reading program affected

[5]Details in Appendix A.

[6]This evaluation is part of a series of experimental studies conducted in collaboration with NewGlobe's Learning Innovation team (Gray-Lobe et al., 2025a, 2024, 2025b). The Learning Innovation unit aims to improve learning outcomes by refining NewGlobe's materials and testing whether variations in materials

outcomes of fifth-graders in the present study cannot be empirically tested.

3.2 Randomization & inclusion criteria

Bridge Nigeria operated 63 schools during this program. One school was excluded from randomization because it operated a slightly different model (known as "Bridge Plus"). The remaining 62 schools were randomly assigned to either the interleaved or blocked problem set condition. Each school contained one grade five classroom. We use school and classroom interchangeably to refer to the unit of randomization.

Randomization was stratified based on pre-assignment characteristics, including lesson completion rates, student-teacher ratios, school urban/rural classification, and whether the school had a grade five class the previous year.[7][8] Figure 2 illustrates final sample construction. Twelve strata were formed in total, with 28 schools assigned to interleaved practice and 34 to blocked practice. Post-randomization, two strata were found to contain only one school each. The randomization procedure deterministically assigned these strata to the comparison group, so these schools were dropped from the analysis.

The analysis sample is restricted to ensure that estimated effects capture the program's causal impact on learning rather than changes in enrollment or selective attrition. Student entry and exit can complicate interpretation.[9] To avoid complications from unobserved entrants, the sample excludes students who enrolled after the intervention began. This restriction simplifies the analysis: new entrants lack pre-intervention test scores and an observable comparison group for attrition. Consequently, estimated impacts pertain to incumbent students. As shown below, including new entrants does not materially change the main results.

The sample is further restricted to students confirmed to be enrolled in Bridge at the time the interleaving program was launched. In many cases, families may withdraw their children from Bridge without giving Bridge notice. Missing test scores can be a strong indicator of whether students are truly enrolled. The sample is therefore restricted to those students who have test score results from the period before the program's start. Consistent with the hypothesis that missing test scores often indicate that students have withdrawn, the sample

are effective enough for large-scale implementation. NewGlobe has also collaborated with other researchers (Schueler and Rodriguez-Segura, 2020; Romero et al., 2022; Esposito Acosta and Sautmann, 2022).

[7]Strata were defined by indicators for schools with lesson completion rates above 75% (approx. the median), student-teacher ratios of ≤ 15, 15–30, or >30, and urban classification.

[8]Urban/rural classifications were based on Bridge's cost-of-living salary adjustments.

[9]A note on where exiting students go: Gray-Lobe et al. (2024) report that over 30 percent of students exited Bridge schools in Kenya over a school year. Table 3 of Gray-Lobe et al. (2022) similarly shows exit rates of about 20 percent among non-scholarship students, with lower attrition among scholarship recipients, suggesting that difficulty paying fees drives much of the turnover. Most exiting students transferred to lower-cost public schools, though some moved to other private schools. Over two years, all primary-school-aged students initially enrolled at Bridge remained enrolled in some primary school by study end.

that satisfies this condition has substantially lower levels of test score attrition in follow-up periods. Given these concerns, students were included if they met two criteria: 1) Their unique identifier appeared in a NewGlobe mathematics test data file from immediately before the program's start and 2) Pre-assignment data included their gender, age, and enrollment date. Based on these criteria, 525 students were excluded. One interleaved school had no eligible students and was dropped. Robustness tests indicate that these inclusion criteria do not affect the main results (see Appendix Table A9).

3.3 Participants

We use student-level and school-level data collected by Bridge.[10] Table 1 shows that both student-and school-level characteristics are similar for the interleaved and blocked groups.

The study sample includes 687 students (Table 1). At baseline, the average age was 9.81 years, and about half were female. Students had been enrolled in a Bridge Nigeria school for 1.23 years on average when the interleaving program started.[11] The previous year's average student-teacher ratio was 23, and teachers completed scheduled lesson guides 77% of the time.

Most schools in the sample are in Lagos state, one of sub-Saharan Africa's most densely populated regions. According to Bridge's classification, 68% are designated as rural, although many of the Lagos schools are located on the urban periphery rather than in remote rural areas. Eight percent are in Osun state, a less urban state, where all schools are classified as rural.

3.4 Measures (Test scores)

NewGlobe developed termly posttests measuring short-term retention and a yearly cumulative assessment measuring long-term retention of material covered in the "my turn/your turn" sessions.[12] Each test was composed of thirty open-ended items graded 'correct' or 'incorrect', with no partial credit. Grade 5 math teachers received test items on tablets and wrote them on the board for students to answer in their exercise books. Academy managers assigned a different teacher within the same school to grade the tests. After the test, students received the correct answers. Teachers report the total number of correct responses

[10] Although Bridge is a for-profit organization, we believe it conducted this experiment to explore the effectiveness of a pedagogical intervention, with results – positive or negative – having no impact on its reputation or profitability.

[11] Enrollment dates reflect the earliest entry into any Bridge Nigeria school, even if a student changed schools.

[12] These tests supplemented Bridge Nigeria's regular mid- and end-term summative assessments.

for each student; item-level data was not recorded. For analysis, raw scores were standardized using the comparison group's mean and standard deviation. Figure 3 shows the test administration timeline.

Posttests (short-term) Each posttest consisted of 30 items designed to measure short-term retention of material taught during the term. Test items were sampled uniformly from the "your turn/my turn" practice exercises and adapted to create parallel assessment items. On average, the interval between a topic's introduction and its corresponding posttest was 39 days for the first posttest, 38 days for the second, and 29 days for the third. The first posttest included three items introduced within five days of testing, while the second and third each included only one such item. More than two-thirds of all posttest items were based on content introduced over ten days prior. The assessments spanned a wide range of mathematical topics that varied across terms, with substantial emphasis on basic arithmetic, fractions, percentages, and geometry (Table 2). We use the phrase "short-term retention" to distinguish posttest outcomes from the cumulative assessment, although we note that the posttests are primarily composed of material that was introduced weeks before the assessment.[13]

Our preferred measure of short-term retention averages a student's *observed* posttests.[14] In both blocked and interleaved classrooms, 97% of students have an average posttest score (Table 3, Panel A, Column 1). Follow-up on individual posttests is lower, with students in interleaved classrooms five percentage points less likely to be observed across all three posttests (Appendix Table A2, Column 2).

Cumulative test score (long-term) The cumulative assessment consisted of 30 items and was administered five days after the interleaving program concluded. It covered material from the entire academic year and was designed to align closely with the prior posttests. Of the 30 items, 30% mirrored questions from the Term 1 posttest, 37% from Term 2, and 33% from Term 3 (Table 2), ensuring balanced representation across terms. On average, the content assessed had been introduced 145 days before the cumulative assessment was

[13]In the interleaving literature, "short-term" often refers to same-day (e.g., Patel et al., 2016) or same-week effects (e.g., Taylor, 2008; Taylor and Rohrer, 2010b). Compared to these, our short-term results reflect longer retention. However, the retention intervals align with other large field experiments on interleaving, and prior studies suggest very short-run impacts are often smaller than longer-run effects (Brunmair and Richter, 2019).

[14]This approach improves conciseness, reduces student-level attrition (as it includes students with at least one posttest result), and enhances statistical power (e.g., Kling et al., 2007). Estimates of the effect on this index can be interpreted approximately as the average of effects on individual posttests. Figure 5 shows similar impacts across all three posttests. The p-value from a joint test of hypothesis that impacts for all terms was zero is 0.099.

administered.

Approximately 78% of students in comparison schools have cumulative test data. Students in interleaved schools are four percentage points less likely to be observed for this test, but the difference is not statistically significant (Panel A, Column 2, Table 3).

Note that the cumulative assessment includes material that was practiced almost nine months earlier as well as material that may have been practiced shortly before the assessment. Effects observed on this assessment may reflect a mixture of both impacts on short- (material from Term 3) and long-term retention (material from Terms 1 and 2). Estimates of the interleaving effect on the cumulative assessment may be biased in the direction of the impact on short-term retention relative to a test that excludes Term 3.

A note on differential attrition Differences in follow-up rates between interleaved and blocked classrooms may raise concerns about selective attrition. Lower follow-up in interleaved classrooms is largely explained by slightly lower attendance on test days (Appendix Table A4). As previously noted, interleaving is a form of desirable difficulty—by design, it makes practice more effortful (e.g., Taylor, 2008; Ziegler and Stern, 2014) – which may have discouraged some students from participating.[15] Such discouragement effects could disproportionately affect students depending on their initial skill levels.

While we cannot rule out selective attrition, the evidence suggests it is unlikely. The effects of interleaved practice on individual posttests are similar over time (Figure 5) while the attrition varies over time. Most students completed at least one test, reducing the risk of bias due to attrition for the short-term measures of retention. Importantly, attrited and non-attrited students show no significant differences in baseline performance (Appendix Table A3).

3.5 Empirical framework

We want to estimate the effect of increasing exposure to desirable difficulties in math practice on test scores $Y_{s,i,j}$ where s indexes outcomes, i students, and j schools. We estimate the following linear model of test scores

$$Y_{ijs} = \alpha_s + \beta_s Z_j + \gamma_s p_j + \epsilon_{ijs} \tag{1}$$

where $Z_j \in 0, 1$ indicates the classroom's assignment to the interleaving group, p_j is the probability that school j would be assigned to the interleaving condition given the school's

[15] Since students were not informed in advance about tests, it is unlikely they deliberately avoided assessments.

randomization stratum, and ϵ_{ijs} is an idiosyncratic error term potentially containing a common classroom component.

Fitting the data to Equation 1 yields an estimate, $\hat{\beta}_s$, of the effect of interleaving versus blocking problem sets. Given random assignment of Z_j, $\hat{\beta}_s$ is an unbiased estimate of the average effect of increasing interleaving in problem sets compared to blocking on outcome s. Because blocked practice was the status quo ex ante, we refer to the blocked condition as a "comparison" group, and we call the estimate $\hat{\beta}_s$, the effect of interleaved relative to blocked practice, the "interleaving effect". Standard errors are clustered at the school level.[16]

In all specifications, we control for the probability that a school was assigned to the interleaving condition rather than including strata fixed effects. Treatment status does not vary within some strata in the estimation samples used in this paper, particularly in specifications examining heterogeneous effects by baseline test scores, since baseline data are missing for some schools. Controlling for the probability of treatment is sufficient to ensure unconfoundedness while preserving sample size. By Theorem 1 of Rosenbaum and Rubin (1983), if $Y(1), Y(0) \perp Z \mid X$, where X denotes the randomization strata, then it also follows that $Y(1), Y(0) \perp Z_i \mid e(X)$, where $e(X)$ is the probability of treatment conditional on a unit's randomization stratum. Results are broadly similar when strata fixed effects are included instead, though less precise due to reduced residual variation in interleaving status (Appendix Table A10).

Motivated by prior literature showing that interleaving may have heterogeneous impacts on students depending on their initial level of mathematical knowledge (Ostrow et al., 2015), we extend the analysis to explore distributional impacts. First, we test whether a student's baseline test score predicts variation in the impact of interleaving on students by estimating the following linear regression model:

$$Y_{ijs} = \delta_s + \kappa_s Z_j + \lambda_s Z_j \times y_i + \mu_s y_i + \pi_s p_{ij} + \eta_{ijs} \tag{2}$$

where y_i is a baseline math test score, which has a mean of zero in the blocked practice

[16]While conventional in school-clustered randomized evaluations, we note that clustering at the randomization unit level may produce misleading inference. Chaisemartin and Ramirez-Cuellar (2020) show that, in cluster-randomized evaluations, when randomization strata contain a small number (less than 10) of randomization units (schools, in this case), clustering at the stratum level produces more accurate inferential error rates. Only two (out of ten) strata contain more than 10 schools. However, because the number of strata is also small, clustering at the strata level can produce downward-biased estimates of standard errors of the interleaving effect (Cameron and Miller, 2015). In our specifications (which do not include randomization strata fixed effects) errors clustered at the school level may produce slightly conservative inference in expectation (Chaisemartin and Ramirez-Cuellar, 2020). Clustering standard errors at the strata level in conjunction with wild cluster bootstrap inference (Cameron and Miller, 2015) does not meaningfully affect our main results (see Appendix Table A5).

group. We use the mean of the three midterm and endterm math scores before the start of the program to construct the baseline math test score. Estimation by OLS yields $\hat{\kappa}_s$, which gives the effect of interleaving on a student with a baseline test score equal to the population mean of zero, and $\hat{\lambda}_s$, the difference in the effect for a student with a baseline test score one standard deviation above the mean. Second, we also use quantile regression to examine distributional impacts.

Estimation of Equation 2 can understate the degree to which the interleaving effect varies with baseline math skills if the scores contain measurement error. Quantile regression gives the interleaving effect, $\hat{\beta}_s^{\tau}$, *at* a chosen percentile τ of the test score distribution. Assuming the relative ranks of students are preserved regardless of the form of practice, quantile regression estimates can be interpreted as describing how treatment affects different parts of the outcome distribution, though they do not directly identify effects on students at fixed initial rankings. The rank preservation assumption is strong, especially given the evidence of larger impacts for lower-performing students. We note the similarity in estimates from Equation 2 as support for the interpretation of quantile regression estimates in this way. As we discuss below, likely violations of rank preservation would tend to mean that our results understate the distributional impacts of interleaving.

3.6 Data availability

All data and study analysis code used in this study are publicly available at the Harvard Dataverse (https://doi.org/10.7910/DVN/BBS0SH).

3.7 Pre-registration

The study was pre-registered internally by NewGlobe before implementation, though these documents were not publicly archived. The pre-registration specified that the intervention aimed to leverage interleaving to improve students' ability to match problems to appropriate solution strategies, and thereby improve performance on cumulative assessments. Impacts were to be evaluated using specially administered posttests and a cumulative end-of-year assessment.

4 Results

This section examines the aggregate effect of interleaving compared to blocked practice on shorter-term posttests, measuring learning over the course of an academic term, as well as on the cumulative assessment at the end of the year. Extensions of the analysis consider (a)

the expected impact on the cumulative assessment given the impact observed on the termly posttests, (b) distributional effects, and (c) the impact on long-term retention of term 1 and term 2 material at the end of the year.

4.1 The aggregate effect of interleaving in the short and long term

Interleaved practice increases short-term retention by 0.28 standard deviations ($p = .022$) (Table 3, Panel B, Column 1), an effect equivalent to moving a student from the median of the comparison group distribution of test scores to the 61st percentile. This effect is similar to the short-term effect of 0.34 standard deviations found in a meta-analysis of evaluations comparing interleaving to blocked practice in mathematics education (Brunmair and Richter, 2019).

Impacts on individual posttests are similar across the three posttests (Figure 5). The largest impact (0.31 standard deviations) comes from the third posttest, the test with the shortest average retention interval (29 days compared to 39 and 38 for the first and second posttests, respectively). However, the difference between the interleaving effect on the third posttest and the other posttests is not statistically significant.

Despite large impacts on each of the short-term posttests, the estimated interleaving effect on cumulative test scores is 0.03 standard deviations ($p = 0.858$) (Table 3, Panel B, Column 2). The null hypothesis that the interleaving effect on the cumulative assessment is zero cannot be rejected. The 95% confidence interval [-0.28, 0.34] indicates that moderate positive or negative effects cannot be ruled out.

The longer-term interleaving effect is smaller than the short-term effect (Table 3, Panel B). We test whether the effect on longer-run retention (the effect on the cumulative assessment) is equal to the short-term retention (the effect on the posttest) by estimating the effects jointly using seemingly unrelated regression (SUR) and then test the hypothesis that the estimated effects on the two outcomes are equal. The p-value from this test is 0.02.

4.2 Distributional effects

Interleaving appears to have larger, more persistent, and more robust impacts on initially lower-achieving students than on higher-achieving students. As discussed below, this pattern of effects is found both in estimates of the interaction between baseline test scores and the interleaving condition as described in Equation 2 and in quantile regression estimates.

For students at the mean of the test score distribution, interleaving increased posttest scores by 0.35 standard deviations (3, Panel C, Column 1). For a student with a baseline test score one standard deviation above the mean, interleaving increased posttest scores by

0.16 standard deviations. Conversely, for students one standard deviation below the mean, interleaving increased posttest scores by 0.54 standard deviations. For students at the mean, the estimated interleaving effect on the cumulative assessment is 0.02 standard deviations, and the estimated interaction term is -0.05 standard deviations (Table 3, Panel C, Column 2). Neither coefficient is statistically distinguishable from zero. The interaction with baseline performance is negative – mirroring the results for the posttest – and the implied impact on a student one standard deviation below the mean is approximately 0.07 standard deviations, though this estimate is not statistically distinguishable from zero for this subgroup.

The effect of interleaved practice was largest at the bottom of the distribution for both the posttests and cumulative assessment. On the short-term test, effects at the 10th, 20th, 25th, and 50th percentiles are respectively 0.49, 0.44, 0.44, and 0.38 standard deviations and statistically significant.[17] The estimated effect on the 20th percentile is 0.49 standard deviations and highly statistically significant (Figure 6) on the long-term test.[18] The estimated effect at the 90th percentile is -0.33 standard deviations, and the null hypothesis can be rejected at the five percent level, suggesting that interleaving may be harmful for top performers.

4.3 Interpretation of the absence of a cumulative effect

This section briefly discusses the interpretation of the absence of an effect on the cumulative assessment. We consider whether the effect reflects students in the blocked condition gaining on those in the interleaving condition or the fading memories in the interleaving group, whether the effect could reflect unit inconsistency between the posttests and the cumulative assessment, and whether the effect could be driven by selective attrition.

4.3.1 Catch up versus fade out

The absence of an effect on the cumulative assessment may reflect either the comparison group learning more rapidly or the interleaving group forgetting. Unfortunately, data limitations – specifically a lack of item-level data – prevent this study from providing a clear answer to this question. While we cannot rule out forgetting, several factors suggest it played a minor role. Short-term assessments measured skills learned over an average of 36 days, reflecting long-term memory consolidation rather than short-term test cramming. The similarity of math topics and frequent testing provided ample repetition and retrieval practice, so it seems unlikely that skills would be so quickly lost. Additionally, students in blocked

[17] For a visual comparison of the test score distributions, see Figure 4.

[18] Appendix Table A7 reports results in table form.

conditions performed better on the cumulative assessment than on the Term 2 or Term 3 assessments (Appendix Figure A1). Seventy percent of the items on the cumulative assessment mirrored Term 2 and 3 posttests.

4.3.2 Variation in content

We find little support for the view that the cumulative assessment either measures a different set of skills. As discussed above, the items in the cumulative assessment were designed to mirror items on the posttest directly.

Nonetheless, the items themselves were different, and the assessments may have weighted specific skills differently due to sampling or idiosyncratic differences between the items.

Variation in the content in the cumulative assessment and posttests could explain the absence of an effect on the cumulative assessment in two ways. First, the cumulative assessment could be measuring a set of skills that were less affected by the interleaving program. Second, the dispersion of performance on the cumulative assessment may reflect a larger difference in learning than that captured on the more narrowly focused posttests. In this view, the smaller estimated effect on the cumulative assessment may be consistent with persistent effects in the underlying skills measured (albeit noisily) by the cumulative assessment.

In this view, large impacts on posttests may not translate into large impacts on the cumulative assessment because the gains on the posttests translate into smaller gains on the cumulative assessment, which either measures different skills or where a standard deviation reflects a larger change in math skills.

To address this question, we estimate the impact on the expected cumulative test score given students' performance on the posttests, where the predicted cumulative assessment performance is given by OLS regression of the cumulative assessment on posttest scores. As an initial step, we estimate the conditional expectation of the cumulative assessment score, $Y_{ij,\text{cumulative}}$, as a linear function of the termly assessments $Y_{ij,\text{Term t}}$ in the blocked group

$$Y_{ij,\text{cumulative}} = \alpha + \sum_{t=1}^{3} \delta_t Y_{ij,\text{Term } t} + \varepsilon_{ij} \ . \tag{3}$$

We then estimate Equation 1 using the prediction from Equation 3, $\hat{Y}_{ij,\text{cumulative}} = \hat{\alpha} + \sum_{t=1}^{3} \hat{\delta}_t Y_{ij,\text{Term } t}$ as the dependent variable. The estimated coefficients $\hat{\delta}_t$ represent the change in the difference in the cumulative score associated with an increase in the termly posttest score.[19]

[19]This analysis inverts the familiar framework of surrogate analysis (Prentice, 1989; Athey et al., 2025). Surrogate analysis is typically used to forecast treatment effects on long-term outcomes when effects on short-term (surrogate) measures, and (b) the relationship between short- and long-term outcomes can be

The estimated effect of interleaving on the predicted score reflects the impact that would have been expected had the impacts on the short-term assessments been observed and the correlation between the short-term and long-term outcomes been known. Estimates of standard errors are obtained by bootstrapping to account for sampling variance in the estimation of the prediction (Athey et al., 2025).

Posttests are highly predictive of the cumulative assessment. The R^2 from the OLS regression in Equation 3 is 0.62. Scores from Term 2 and Term 3 appear to be most predictive, with coefficients of 0.35 and 0.39 respectively (Appendix Figure A2). Term 1 performance is less predictive with a coefficient of 0.17.

The relationship between posttest and cumulative-assessment performance implies that the observed posttest gains should have translated into substantially larger improvements on the cumulative assessment than were actually observed. The estimated aggregate effect expected, given the short-term gains, was 0.24 standard deviations, nearly ten times the effect, 0.028 standard deviations, observed on the observed cumulative assessment (Panel A of Figure A3). The gap is especially striking when considering initially lower-performing students. Panel B of Figure A3 also reports the impacts using both the predicted and the actual cumulative score on those students one standard deviation below the mean of the sample. The effect that would have been expected, given the short-term impacts on lower-performing students, was 0.53 standard deviations. However, the estimated effect for these students was only 0.07 standard deviations.

4.3.3 Attrition

We find little evidence that the cumulative assessment might be biased downward due to selective attrition. The risk that the interleaving effect for the cumulative assessment is compromised by selective attrition is greater than that for the posttest index, for which follow-up is high. Appendix Table A3 reports covariate balance between the interleaved and blocked groups conditional on follow-up on all individual posttests and the cumulative assessment. On observable characteristics, we see no evidence of differences in the sample composition conditional on follow-up for any single test. Also, despite the similar follow-up rates on the Term 3 and cumulative assessments (Appendix Table A3), we find very large

estimated, often using external data. Step (b) provides parameters describing how long-term outcomes respond, on average, to changes in short-term outcomes. Combined with the estimated short-term effects, these parameters generate a forecast of the long-term impact. For this procedure to yield unbiased predictions, two key assumptions are required: (i) the estimated relationship between short- and long-term outcomes reflects the true average causal response, and (ii) the treatment affects long-term outcomes only through its effects on the short-term measures (the surrogacy assumption). In the present application, both short- and long-term effects have been observed, and we are evaluating the performance of the posttests as surrogates for the cumulative assessment.

impacts on the Term 3 posttest (Figure 5).

Appendix Table A11 reports Lee (2009) bounds for the estimates of the effect on the posttest and cumulative assessment. Lee bounds address selective attrition concerns by computing the range of treatment effects consistent with the observed data under the extreme selectivity assumptions. The upper bound assumes that the students missing from the interleaved group are those who would have scored highest, so that the estimate in Table 3 understates the effect. The lower bound assumes the opposite — that the missing students in the interleaving group would have scored the lowest. Given the absence of evidence of selective attrition on baseline test scores, we view Lee bounds as giving a conservatively large range of estimates. Even so, the upper bound of the estimate is 0.139 standard deviations, well below the effect on the posttests.

4.4 Gender heterogeneity

We tested for and found no evidence of heterogeneous impacts by gender. Results are reported in Appendix Table A6.

4.5 Cost-effectiveness

We briefly discuss the cost-effectiveness of the interleaving program. This exercise should be interpreted cautiously because cost-effectiveness depends critically on which outcome is used. The interleaving program increased performance on the termly posttests by 0.28 standard deviations, but the estimated effect on the cumulative end-of-year assessment was only 0.03 standard deviations and statistically indistinguishable from zero. At this stage, we view the cumulative assessment results as uncertain evidence on the long-run achievement effects of interleaving. We therefore focus on the posttest impacts and interpret the calculations below as answering the following question: how cost-effective would the program be if the learning gains observed on the termly posttests could be sustained?

The costs of the program are primarily the costs of modifying practice materials and informing teachers. The intervention did not require additional teachers, instructional time, student materials, or classroom infrastructure. We assume conservatively that rearranging practice items requires 10 percent of the annual time of an entry-level staff member earning $40,000 per year, or $4,000. We further assume that informing teachers requires one half-day of staff time per school. With 62 schools and an assumed 250 working days per year, this adds $4,960. Under these assumptions, total implementation costs are $8,960. Dividing by the 687 students in the analysis sample gives a cost of $13.04 per student. This is likely conservative for implementation at scale because the cost of revising centrally prepared practice materials

is largely fixed and would be spread over more students in larger systems.

Following the convention in the education cost-effectiveness literature of converting 0.8 standard deviations into one Learning-Adjusted Year of Schooling, the 0.28 standard deviation posttest effect corresponds to 0.35 Learning-Adjusted Years of Schooling per student. At a cost of $13.04 per student, this implies approximately 2.7 Learning-Adjusted Years of Schooling per $100. This calculation is substantially less favorable if the cumulative assessment point estimate is used instead. The cumulative effect of 0.03 standard deviations corresponds to 0.04 Learning-Adjusted Years of Schooling per student, or approximately 0.29 Learning-Adjusted Years of Schooling per $100.

These calculations should be viewed as illustrative rather than definitive. They highlight that interleaving may be inexpensive to implement in settings where instructional materials are centrally produced, but they also underscore that the welfare interpretation of the intervention depends on the outcome of interest. If the relevant objective is to improve performance on material practiced during the term, interleaving appears highly cost-effective. If the relevant objective is end-of-year cumulative achievement under fixed curriculum coverage, the present estimates do not provide evidence of comparable cost-effectiveness.

5 A Simple Model of Retrieval Difficulty and Learning

We present a model in which the effectiveness of practice depends on the retrieval difficulty experienced by the learner. The model has four implications which are consistent with the empirical results. First, increasing difficulty can raise learning when practice would otherwise be too easy. Second, the effect of doing so depends on students' existing knowledge so that the impacts of increasing difficulty are inherently heterogeneous. Third, an initial learning advantage fades when students subsequently face common practice opportunities.

The model, however, has two implications that suggest that introducing initial desirable difficulties can still be useful even when effects fade out. First, introducing desirable difficulties initially can reduce the effort or adversity that students face in reaching a particular level of understanding of a topic. This may have important implications if students can become discouraged or develop their academic self-image. Second, extending the model to allow for practice of multiple skills, increasing desirable difficulties initially can allow for more rapid accumulation of knowledge, so combining changes in the difficulty of initial practice with changes in the pace of instruction could sustain impacts.

The model considers only shocks to the initial difficulty of practice in a context where students face repeated opportunities for effortful practice, which we view as largely capturing the setting of our study. Modifying the difficulty of subsequent difficulty could also sustain

effects.

Consider a student with $T+1$ opportunities to practice a skill in periods $t \in \{0, \ldots, T\}$. Let K_t denote the student's knowledge at the beginning of period t. Knowledge evolves according to

$$K_{t+1} = (1-\delta)K_t + g(q_t), \tag{4}$$

where $\delta \in (0,1)$ is the rate of forgetting and $g(q_t)$ is expected learning from the practice opportunity.

Let

$$q_t = q(d_t, K_t)$$

denote retrieval difficulty, where d_t is the absolute difficulty created by the design of practice. We assume that $q(d,K)$ is continuously differentiable and satisfies

$$q_d(d,K) > 0 \qquad \text{and} \qquad q_K(d,K) < 0.$$

Thus, more demanding tasks are more difficult, while the same task is easier for a student with greater existing knowledge.

Expected learning is

$$g(q) = b(q)p(q),$$

where $b(q)$ is the learning benefit conditional on successful retrieval and $p(q)$ is the probability that retrieval succeeds. We assume

$$b(q) > 0, \qquad b'(q) > 0,$$

and

$$p(q) \in (0,1), \qquad p'(q) < 0.$$

More difficult retrieval, therefore, generates a larger benefit when successful but is less likely to succeed. We use this formulation to represent the central mechanism emphasized by the desirable-difficulties framework: more difficult retrieval is more productive conditional on success, but becomes less pedagogically valuable as the probability of successful retrieval falls (e.g. Bjork and Bjork, 2011; Pyc and Rawson, 2009).

Define the proportional rates of change

$$\eta_b(q) = \frac{b'(q)}{b(q)} \qquad \text{and} \qquad \eta_p(q) = -\frac{p'(q)}{p(q)}.$$

We assume that $\eta_b(q)$ is weakly decreasing, that $\eta_p(q)$ is weakly increasing, and that they

cross exactly once. Specifically, there exists a unique q^* such that

$$\eta_b(q) > \eta_p(q) \quad \text{for } q < q^*,$$

and

$$\eta_b(q) < \eta_p(q) \quad \text{for } q > q^*.$$

Because

$$g'(q) = b(q)p(q)\left[\eta_b(q) - \eta_p(q)\right], \tag{5}$$

expected learning is increasing for $q < q^*$ and decreasing for $q > q^*$. Practice that is too easy generates little learning, while practice that is too difficult is unlikely to result in successful retrieval, similar to the idea of a Zone of Proximal Development (Vygotsky, 1978).

Fadeout over time. The key dynamic feature of the model is that learning from one practice opportunity changes the difficulty of later opportunities. Consider an intervention that increases practice difficulty in period t from d'_t to $d''_t > d'_t$. If

$$g\left(q(d''_t, K_t)\right) > g\left(q(d'_t, K_t)\right),$$

then Equation 4 implies

$$K_{t+1}(d''_t) > K_{t+1}(d'_t).$$

Because $q_K(d, K) < 0$, the student who learned more subsequently experiences the same practice opportunity as less difficult. The proposition below establishes conditions under which fadeout over time will occur.

Proposition 5.1. ***Fadeout under common subsequent practice.*** *Suppose the assumptions above hold and there exists a constant $\rho \in [0, 1)$ such that, for every relevant d and K,*

$$0 \leq 1 - \delta + q_K(d, K)b(q)p(q)\left[\eta_b(q) - \eta_p(q)\right] \leq \rho, \qquad q = q(d, K).$$

If two students subsequently face the same sequence of practice opportunities, any difference in their knowledge converges to zero without their rankings reversing.

All proofs are collected in Appendix B.

Proposition 5.2. ***Distributional effects.*** *Suppose expected learning is maximized at the unique retrieval difficulty q^*. Then:*

1. *if $q(d''_0, K_0) \leq q^*$, then $\Delta_1(K_0) > 0$;*

2. *if* $q(d_0', K_0) \geq q^*$, *then* $\Delta_1(K_0) < 0$; *and*

3. *if*

$$q(d_0', K_0) < q^* < q(d_0'', K_0),$$

the sign of $\Delta_1(K_0)$ *is ambiguous.*

Cumulative adversity and effort. The model also implies that productive early difficulty can reduce the cumulative burden of learning. One measure of this burden is the expected number of failed retrieval attempts. Since $p(q)$ is the probability of successful retrieval, the probability of failing a practice opportunity is $1 - p(q)$. A second measure is the effort required to complete a practice opportunity. Let $e(q)$ denote this effort, with $e'(q) \geq 0$. Both cases can be represented by a generic burden function $h(q)$ satisfying $h(q) \geq 0$ and $h'(q) > 0$. For failed retrievals, $h(q) = 1 - p(q)$; for effort, $h(q) = e(q)$.

For a mastery threshold $\bar{K}$, define cumulative burden before mastery as

$$B_h(\bar{K}) = \sum_{t=0}^{\tau(\bar{K})-1} h(q_t),$$

where

$$\tau(\bar{K}) = \min\{t : K_t \geq \bar{K}\}.$$

Proposition 5.3. ***Cumulative burden.*** *Consider two regimes that differ only in the difficulty of the initial practice opportunity, with* $d_0^H > d_0^L$, *and that face the same sequence of practice opportunities thereafter. Suppose the initial increase in difficulty is productive, and define*

$$\Delta K_1 = K_1^H - K_1^L = g(q_0^H) - g(q_0^L) > 0.$$

Let $N = \tau_L(\bar{K}) - 1$ *denote the number of subsequent practice opportunities before regime* L *reaches mastery. Suppose that, over the relevant range of future knowledge levels and practice difficulties,*

$$F_d'(K) \geq \lambda \geq 0$$

and

$$-h'(q(d, K))q_K(d, K) \geq \kappa_h > 0.$$

If

$$h(q_0^H) - h(q_0^L) < \kappa_h \Delta K_1 \sum_{s=0}^{N-1} \lambda^s,$$

then

$$B_h^H(\bar{K}) < B_h^L(\bar{K}).$$

Instructional pace. Faster mastery matters when instructional time can be reallocated to additional material. To make this implication explicit, suppose students must master J skills and that instruction moves to a new skill once the current skill reaches the threshold $\bar{K}$. Let τ_j denote the number of practice opportunities required to master skill j. The number of skills mastered after T practice opportunities is

$$M(T) = \max\left\{ m : \sum_{j=1}^{m} \tau_j \leq T \right\}.$$

Proposition 5.4. ***Instructional pace.*** *Suppose productive desirable difficulties reduce the number of practice opportunities required to master each affected skill, so that*

$$\tau_j^H \leq \tau_j^L$$

for every affected skill j. If practice opportunities saved after mastery are reallocated to subsequent skills, then

$$M_H(T) \geq M_L(T)$$

for any fixed amount of instructional time T. Equivalently, regime H requires weakly less time to master any fixed collection of skills.

The proposition clarifies why the pace of instruction is central to the long-run value of desirable difficulties. If curriculum coverage remains fixed, students who learned less initially may eventually catch up through continued practice of the same material. If instruction instead advances when students reach mastery, more productive practice can allow students to master more skills within the same amount of instructional time. Persistent achievement gains therefore require not only faster learning, but also an instructional system that responds to faster mastery by reallocating time toward additional or more advanced material.

6 Conclusion

This paper evaluates a practitioner-designed program that incorporated interleaved mathematics practice into regular instruction over a full academic year. Relative to blocked practice, interleaving changed the organization of practice while holding fixed the curricular content, instructional time, and number of practice problems. Interleaving increased perfor-

mance on termly mathematics assessments by 0.28 standard deviations, with larger gains among initially lower-achieving students. These effects are similar in magnitude to those found in shorter and more tightly controlled studies of interleaving, showing that practitioners can operationalize this insight from cognitive science through ordinary instructional materials.

The termly gains did not, however, translate into a detectable improvement on the cumulative end-of-year assessment. The estimated cumulative effect was 0.03 standard deviations and was statistically distinguishable from the termly effect, although the confidence interval does not rule out moderate positive or negative cumulative effects. Moreover, the relationship between termly and cumulative performance in the comparison group implies that the observed termly gains would ordinarily have predicted a substantially larger cumulative effect. Thus, the results indicate that improvements measured over the course of a term need not accumulate into proportionate differences in achievement at the end of the academic year.

The experiment cannot determine why the termly advantage diminished. The pattern could reflect forgetting among students assigned to interleaved practice, catch-up among students assigned to blocked practice, differences between the assessments, or behavioral responses by students and teachers. The lack of item-level data prevents us from tracing the retention of particular skills or separating these mechanisms. The evidence should therefore not be interpreted as establishing that catch-up caused the smaller cumulative effect.

The model provides one possible explanation based on the mechanism underlying desirable difficulties. Expected learning is greatest at an intermediate level of retrieval difficulty. When more difficult initial practice raises learning, students enter later practice opportunities with greater knowledge. Those later opportunities are consequently easier and may produce less additional learning. Under the stability conditions stated in Proposition 5.1, an initial knowledge advantage necessarily narrows when students subsequently face the same sequence of practice opportunities. This mechanism does not require treated students to lose the knowledge they acquired. Instead, students who learned less initially may benefit more from later opportunities because those opportunities remain more difficult and productive for them.

The model also clarifies why convergence in measured achievement need not imply that more productive practice had no value. When desirable difficulties raise initial learning, they reduce the difficulty of subsequent practice and can reduce the number of practice opportunities required to reach mastery. These benefits are not generally captured by evaluations that compare achievement at fixed points in time. Faster mastery may matter for students' effort, motivation, and use of instructional time even when other students eventually reach the same

level of knowledge. At the same time, making practice excessively difficult can reduce learning, delay mastery, and harm students for whom retrieval is already sufficiently demanding. The appropriate level of difficulty therefore depends on students' existing knowledge.

The long-run effect of more productive practice also depends on how instructional systems respond to faster learning. If all students continue to practice a fixed collection of skills, students who learned less initially have opportunities to catch up. If instruction instead advances when students reach mastery, reducing the practice required for one skill can free time for additional or more advanced material. Interleaving may therefore be most effective as a complement to adjustments in instructional pace rather than as an isolated change to the organization of practice.

These findings suggest two priorities for future research. First, evaluations should measure the histories of individual skills and practice opportunities, allowing researchers to distinguish forgetting from catch-up and to estimate how prior learning changes the productivity of later instruction. Second, experiments should test interleaving jointly with systems that adapt curriculum coverage or instructional pace in response to mastery. Such designs could determine whether faster initial learning can be converted into persistent gains in the breadth or level of material students master.

More broadly, the study identifies the organization of practice as a potentially low-cost margin of educational policy. Interleaving was implemented by practitioners through existing teaching guides across a multisite school network, without additional instructional time or researcher-led delivery. The results show that changing this default can produce substantial gains over meaningful time horizons. They also show that the ultimate effect of more productive practice depends on the difficulty of subsequent learning opportunities and on whether schools use faster mastery to advance instruction.

References

Aitken, C., G. Gray-Lobe, M. Kremer, M. Joshi, and J. de Laat (2025): "Hard to Read: The Impact of Advanced Reading Assignments on Language and Literacy Outcomes," Working Paper.

Akyeampong, K., T. Andrabi, A. Banerjee, R. Banerji, S. Dynarski, R. Glennerster, S. Grantham-McGregor, K. Muralidharan, B. Piper, S. Ruto, J. Saavedra, S. Schmelkes, and H. Yoshikawa (2023): "2023 Cost-Effective Approaches to Improve Global Learning: What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries?" London, Washington D.C., New York: FCDO, the World Bank, UNICEF, and USAID.

Al-Ubaydli, O., J. A. List, and D. Suskind (2019): "The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments," Working Paper 25848, National Bureau of Economic Research.

Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, F. H. Rogers, and S. Sabarwal (2020): *How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions using the New Learning-Adjusted Years of Schooling Metric*, The World Bank.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2025): "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *The Review of Economic Studies*, rdaf087.

Bailey, D. H., G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager (2020): "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions," *Psychological Science in the Public Interest*, 21, 55–97.

Bailey, D. H., G. J. Duncan, C. L. Odgers, and W. Yu (2017): "Persistence and Fadeout in the Impacts of Child and Adolescent Interventions," *Journal of Research on Educational Effectiveness*, 10, 7–39.

Bjork, E. L. and R. A. Bjork (2011): "Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties to Enhance Learning," in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, ed. by M. A. Gernsbacher, R. W. Pew, L. M. Hough, and J. R. Pomerantz, New York: Worth Publishers, 56–64.

Bjork, R. A. (1994): "Memory and metamemory considerations in the training of human beings," in *Metacognition: Knowing about knowing*, ed. by J. Metcalfe and A. P. Shimamura, Cambridge, MA: MIT Press, 185–205.

Brown, P. C., H. L. Roediger III, and M. A. McDaniel (2014): *Make It Stick: The Science of Learning*, Cambridge, MA: The Belknap Press of Harvard University Press.

Brunmair, M. and T. Richter (2019): "Similarity matters: A meta-analysis of interleaved learning and its moderators." *Psychological Bulletin*, 145, 1029.

Cameron, A. C. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 317–372.

Carneiro, P. and J. J. Heckman (2003): "Human Capital Policy," in *Inequality in America: What Role for Human Capital Policies?*, ed. by J. J. Heckman, A. B. Krueger, and B. M. Friedman, Cambridge, MA: MIT Press, 77–239.

Chaisemartin, C. and J. Ramirez-Cuellar (2020): "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" Working Paper 27609, National Bureau of Economic Research.

Claessens, A., M. Engel, and F. C. Curran (2014): "Academic Content, Student Learning, and the Persistence of Preschool Effects," *American Educational Research Journal*, 51, 403–434.

Cunha, F. and J. J. Heckman (2007): "The Technology of Skill Formation," *American Economic Review*, 97, 31–47.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Dillon, M. R., H. Kannan, J. T. Dean, E. S. Spelke, and E. Duflo (2017): "Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics," *Science*, 357, 47–55.

Dunlosky, J., K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham (2013): "Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology," *Psychological Science in the Public Interest*, 14, 4–58, pMID: 26173288.

Esposito Acosta, B. N. and A. Sautmann (2022): "Adaptive Experiments for Policy Choice : Phone Calls for Home Reading in Kenya," Policy Research Working Paper Series 10098, The World Bank.

Evans, D. K. and F. Yuan (2022): "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis*, 44, 532–540.

Gray-Lobe, G., R. Jain, M. Kremer, J. de Laat, P. Schöpfer, and B. Scurlock (2025a): "Does Extra Exam Practice Raise Test Scores? Experimental Evidence from Kenya and Nigeria," Working Paper.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. W. Ozier (2022): "Can education be standardized? Evidence from Kenya," *BFI Working Paper No. 2022-68.* Becker Friedman Institute for Economics, University of Chicago.

Gray-Lobe, G., M. Kremer, J. de Laat, O. Mbonu, and C. Scanlon (2024): "Nudging Parents out the Door: The Impacts of Parental Encouragement on School Choice and Test Scores," Working Paper.

Gray-Lobe, G., M. Kremer, J. de Laat, and W. Wong (2025b): "Details, Delegation, and Control: The Impact of Streamlining Instructions to Teachers on Student Test Scores," Working Paper.

Karpicke, J. D. and H. L. Roediger (2008): "The Critical Importance of Retrieval for Learning," *Science*, 319, 966–968.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75, 83–119.

Kornell, N. and R. A. Bjork (2008): "Learning concepts and categories: Is spacing the "enemy of induction"?" *Psychological Science*, 19, 585–592.

Lee, D. S. (2009): "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," *The Review of Economic Studies*, 76, 1071–1102.

List, J. A. (2022): *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, Penguin Random House.

Mayfield, K. H. and P. N. Chase (2002): "The effects of cumulative practice on mathematics problem solving," *Journal of Applied Behavior Analysis*, 35, 105–123.

Nemeth, L., K. Werker, J. Arend, and F. Lipowsky (2021): "Fostering the acquisition of subtraction strategies with interleaved practice: An intervention study with German third graders," *Learning and Instruction*, 71, 101354.

Ostrow, K., N. Heffernan, C. Heffernan, and Z. Peterson (2015): "Blocking vs. interleaving: Examining single-session effects within middle school math homework," in *International conference on artificial intelligence in education*, Springer, 338–347.

Patel, R., R. Liu, and K. Koedinger (2016): "When to Block versus Interleave Practice? Evidence Against Teaching Fraction Addition before Fraction Multiplication," *Cognitive Science*.

Prentice, R. L. (1989): "Surrogate endpoints in clinical trials: definition and operational criteria," *Statistics in Medicine*, 8, 431–440.

Pyc, M. A. and K. A. Rawson (2009): "Testing the Retrieval Effort Hypothesis: Does Greater Difficulty Correctly Recalling Information Lead to Higher Levels of Memory?" *Journal of Memory and Language*, 60, 437–447.

Rau, M. A., V. Aleven, and N. Rummel (2010): "Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions," in *International conference on intelligent tutoring systems*, Springer, 413–422.

Rau, M. A., N. Rummel, V. Aleven, L. Pacilio, and Z. Tunc-Pekkan (2012): "How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions," *The future of learning: Proceedings of the 10th International Conference of the Learning Sciences*, 64–71, sydney: International Society of Learning Sciences.

Rickard, T. C., J. S.-H. Lau, and H. Pashler (2008): "Spacing and the transition from calculation to retrieval," *Psychonomic Bulletin & Review*, 15, 656–661.

Roediger, H. L. and M. A. Pyc (2012): "Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice," *Journal of Applied Research in Memory and Cognition*, 1, 242–248.

Rohrer, D. (2017): "Interleaved Mathematics Practice Guide," Accessed: 2025-02-11.

Rohrer, D., R. Dedrick, and M. Hartwig (2020a): "The Scarcity of Interleaved Practice in Mathematics Textbooks," *Educational Psychology Review*, 32.

Rohrer, D., R. F. Dedrick, and K. Burgess (2014): "The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems," *Psychonomic Bulletin & Review*, 21, 1323–1330.

Rohrer, D., R. F. Dedrick, M. K. Hartwig, and C.-N. Cheung (2020b): "A randomized controlled trial of interleaved mathematics practice," *Journal of Educational Psychology*, 112, 40.

Rohrer, D., R. F. Dedrick, and S. Stershic (2015): "Interleaved practice improves mathematics learning." *Journal of Educational Psychology*, 107, 900.

Romero, M., L. Chen, and N. Magari (2022): "Cross-Age Tutoring: Experimental Evidence from Kenya," *Economic Development and Cultural Change*, 70, 1133–1157.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Schueler, B. and D. Rodriguez-Segura (2020): "Can Camp Get You Into a Good Secondary School? A Field Experiment of Targeted Instruction in Kenya," Tech. Rep. 197, Annenberg Institute at Brown University.

Taylor, K. and D. Rohrer (2010a): "The effects of interleaved practice," *Applied Cognitive Psychology*, 24, 837–848.

——— (2010b): "The effects of interleaved practice," *Applied Cognitive Psychology*, 24, 837–848.

Taylor, K. M. (2008): *The benefits of interleaving different kinds of mathematics practice problems*, University of South Florida.

Vygotsky, L. S. (1978): "Mind in society: Development of higher psychological processes," .

Ziegler, E. and E. Stern (2014): "Delayed benefits of learning elementary algebraic transformations through contrasted comparisons," *Learning and Instruction*, 33, 131–146.

——— (2016): "Consistent advantages of contrasted comparisons: Algebra learning under direct instruction," *Learning and Instruction*, 41, 41–51.

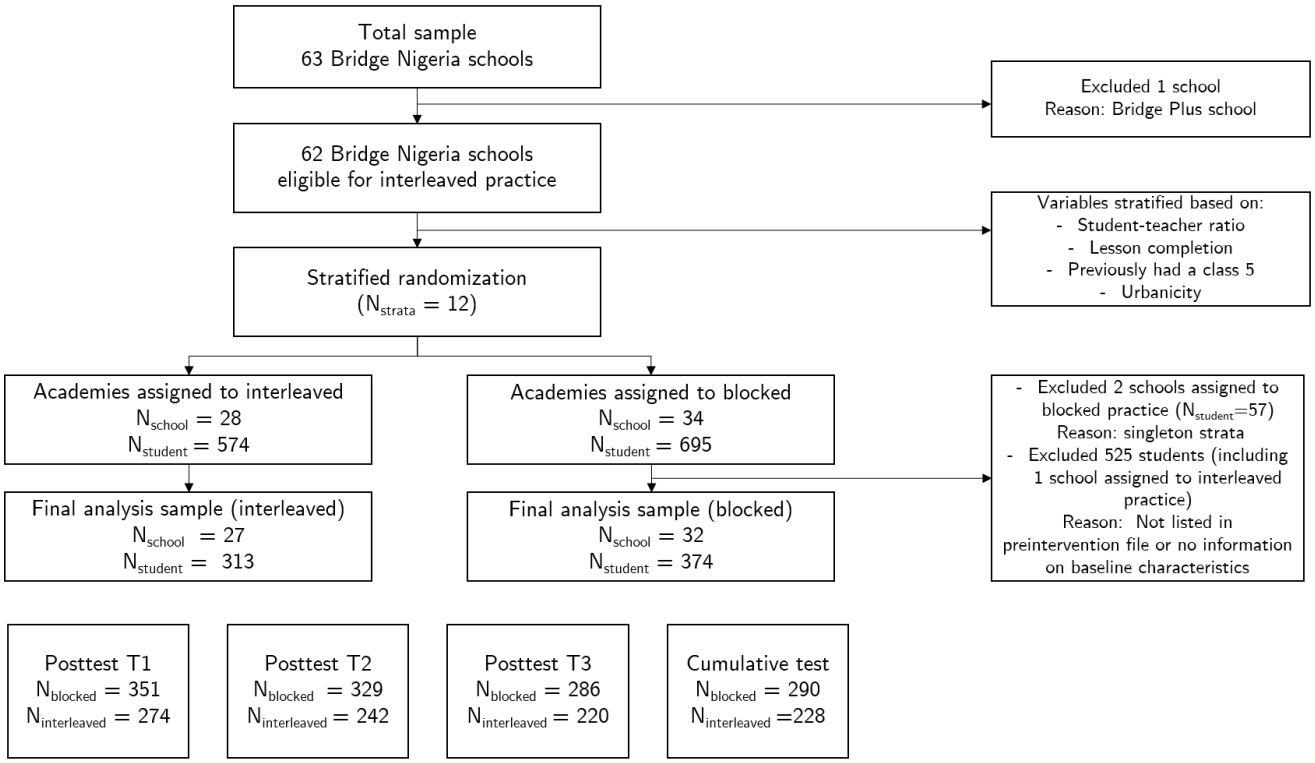
Figures & Tables

Figure 1: Example practice set for blocked and interleaved practice

Blocked practice	Interleaved practice
Write as a percent:	
1) 5/10	1) Write 5/10 as a percent.
2) 7/10	2) Express 65% as a fraction.
3) 8/10	3) Write 8/10 as a percent
4) 7/20	4) List all the factors of 120.
5) 9/20	5) Write 9/20 as a percent.
6) 8/25	6) Write the multiples of 7 between 15 and 50.
7) 9/25	7) The numerator of a fraction is the prime number between 8 and 12. The denominator is a multiple of 10 between 18 and 22. Write this fraction as a percent.
8) 9/50	8) Sam got 9 marks out of 20 in an exam. What is his marks in percent?

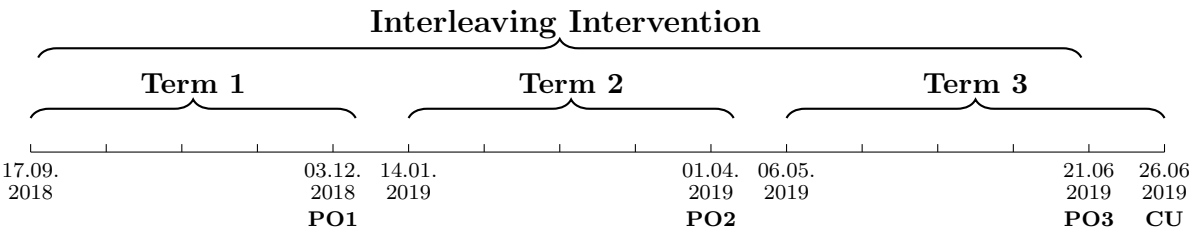
Notes: Examples of blocked practice vs. interleaved practice (independent study). All items in the blocked practice set cover the topic of the day. Items in the interleaved practice set cover the topic of the day (problem 1, 3, 5, 8), topics from previous lessons (problem 2, 4, 6) and an integrated problem (7).

Figure 2: Consort diagram



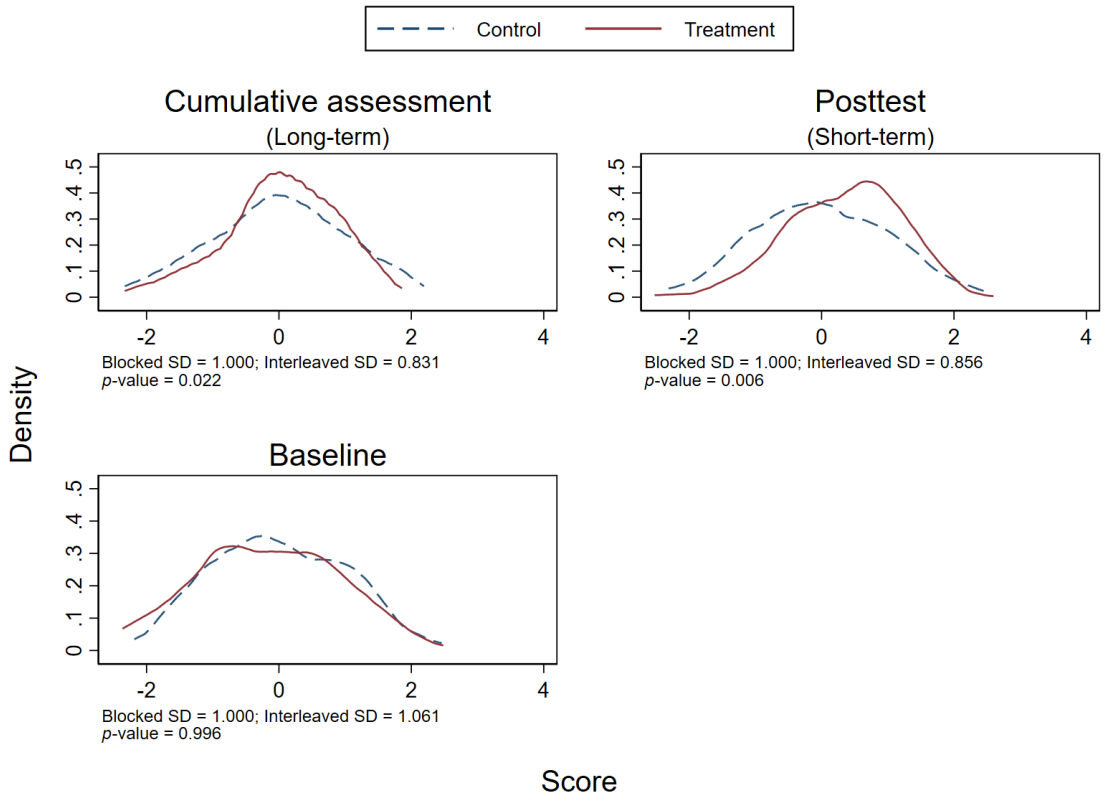
Notes: Figure shows the process of arriving at the final sample of schools and students for analysis. The final row shows the number of students in blocked and interleaved schools that have a test for each of the tests.

Figure 3: Timeline of events



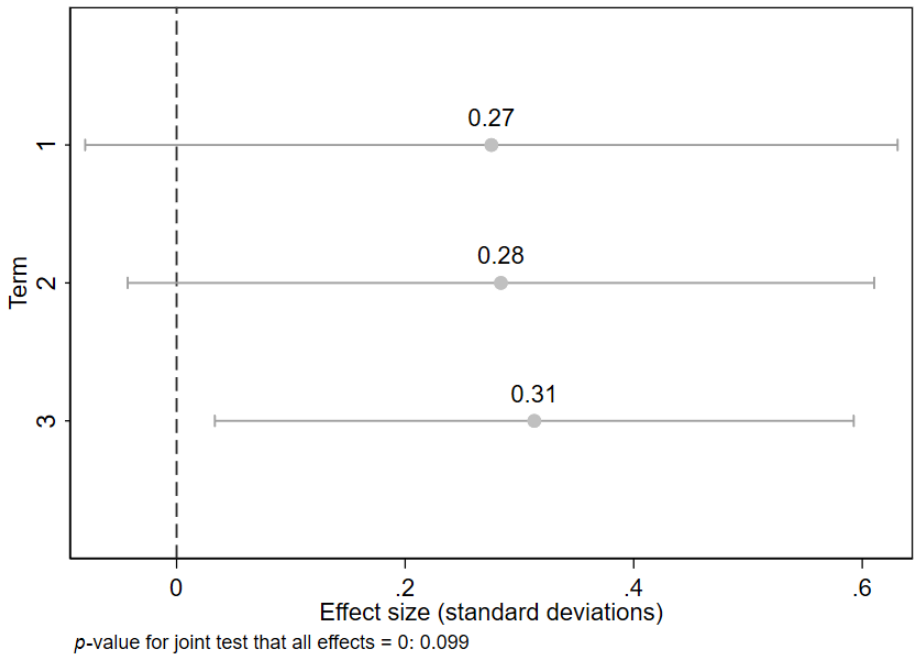
Notes: Figure shows the project timeline. PO=Posttest and CU=Cumulative assessment.

Figure 4: Test score distributions



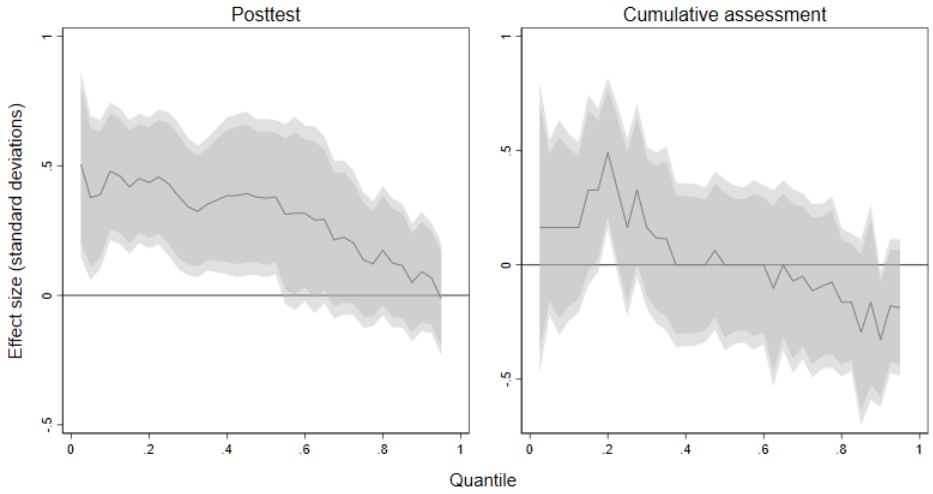
Notes: Each figure plots the kernel density of test scores in blocked and interleaved classrooms. The p-value of the test of equal dispersion comes from the coefficient on a regression of the form in Equation 1 with the absolute deviations from the experimental condition mean as the dependent variable.

Figure 5: Effect on individual posttests



Notes: Coefficient plot of impacts on individual posttests. Error bars indicate 95% confidence interval.

Figure 6: Effects across quantiles



Notes: The figure shows the effects of interleaving across percentiles of the test score distribution. The line represents the point estimate at each percentile. The darker area represents the 10% confidence interval and the lighter area represents the 5% confidence interval.

Figure 7: Zone of proximal development

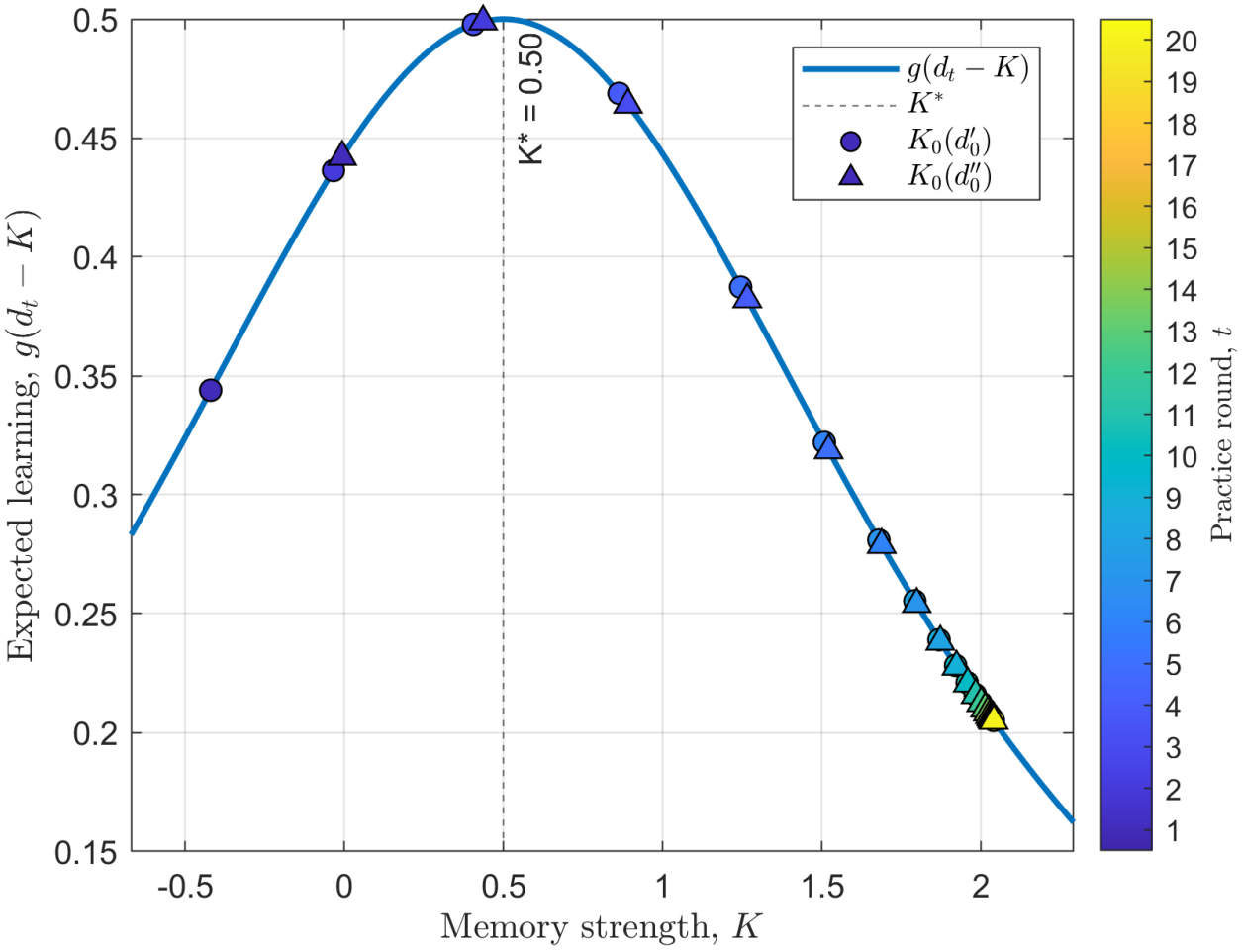
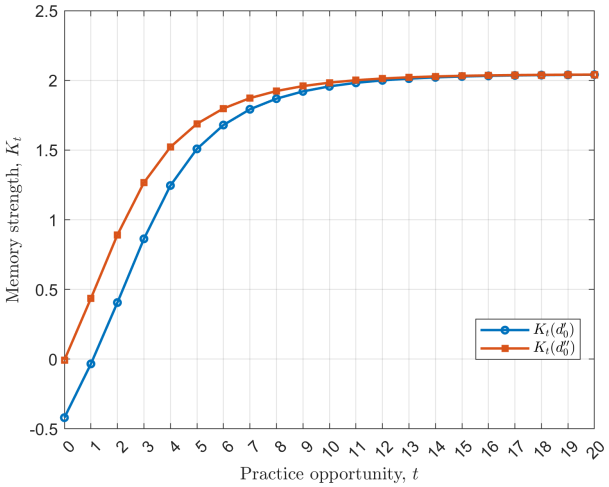
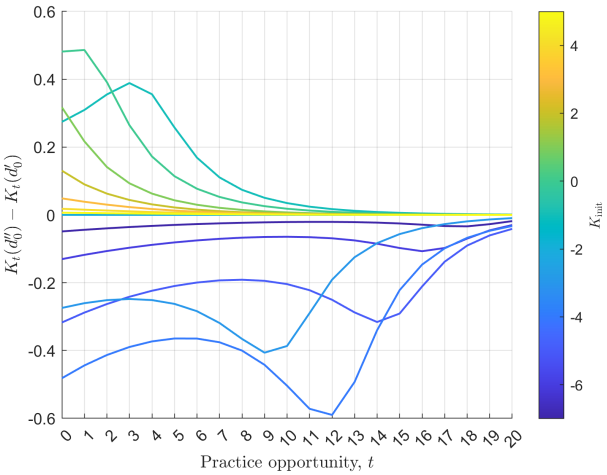


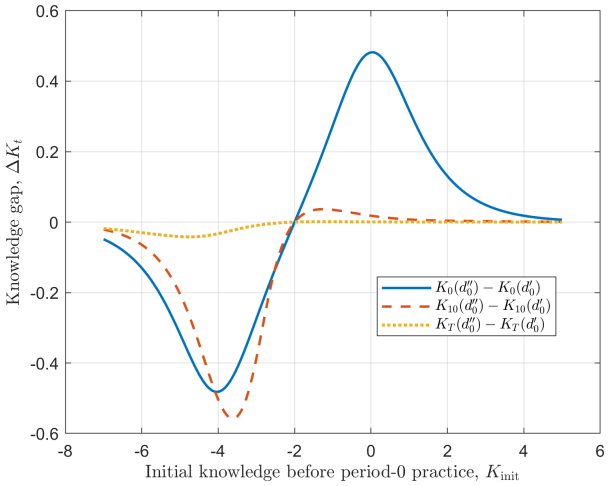
Figure 8: Implications of more difficult initial practice



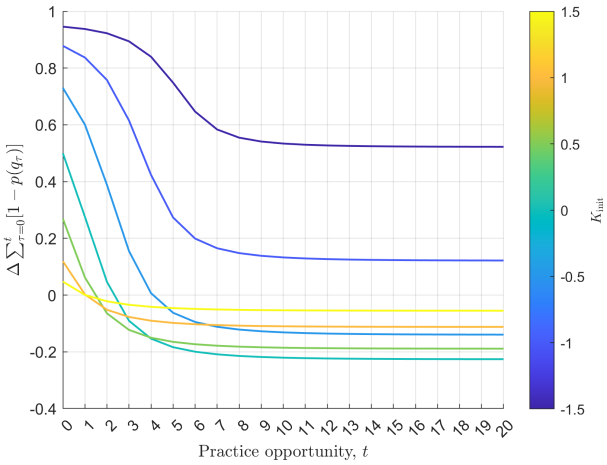
(a) Knowledge accumulation by $t = 0$ practice difficulty



(b) Effect of harder $t = 0$ practice by initial knowledge K_0



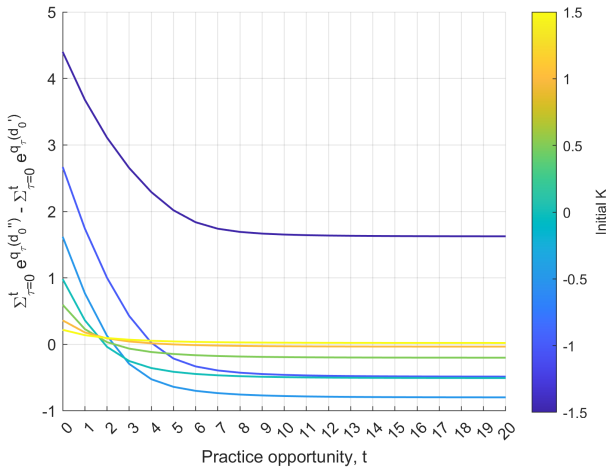
(c) Short-, medium-, and long-run effects by initial knowledge K_0



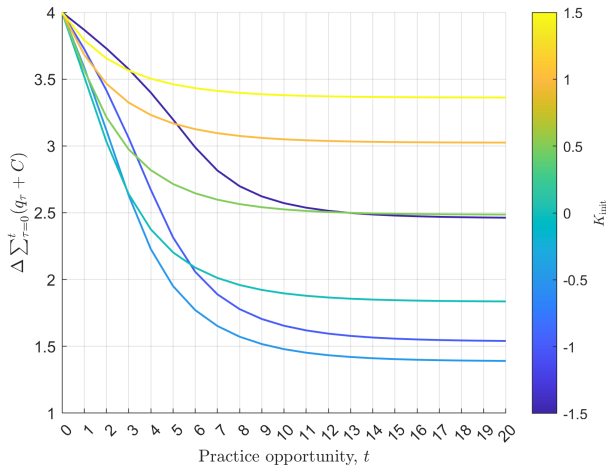
(d) Effect on cumulative failed retrievals

Notes: These figures show the impacts of more difficult initial practice on learning from the parameterized model where $g(q) = e^{\beta q}\frac{1}{1+\exp(\alpha q)}$, with $\alpha = 2$, $\beta = 1$, and the rate of forgetting, $\delta = 0.1$. Panel (a) shows the evolution of knowledge under different levels of initial practice difficulty for $K_0 = 0$. Panel (b) shows how the effect of more difficult initial practice varies with initial knowledge. Panel (c) compares short-, medium-, and long-run effects across the initial knowledge distribution. Panel (d) shows the effect on cumulative difficulty that students incur over time.

Figure 9: Implications of more difficult initial practice on cumulative effort under different effort functions



(a) Effect on cumulative effort ($c(q) = e^q$)



(b) Effect on cumulative difficulty ($c(q) = q + C$)

Notes: The table shows the impact of an initial shock to the level of difficulty on cumulative effort under different functional forms of effort. In panel (a) the effort from each practice round is exponential in difficulty. In panel (b) effort is linear in difficulty.

Table 1: Sample characteristics at baseline

Variable	(1) Blocked N	(1) Blocked Mean/(SE)	(2) Interleaved N	(2) Interleaved Mean/(SE)	(1)-(2) Pairwise t-test N	(1)-(2) Pairwise t-test P-value
Panel A: Student-level characteristics						
Age	371	9.79	312	9.83	683	0.69
		(0.07)		(0.05)		
Years enrolled	371	1.21	312	1.28	683	0.63
		(0.08)		(0.10)		
Female	371	0.48	312	0.54	683	0.11
		(0.03)		(0.03)		
Baseline score	263	0.07	213	-0.05	476	0.41
		(0.11)		(0.13)		
Panel B: School-level characteristics						
Years the school is open	32	1.46	27	1.59	59	0.46
		(0.08)		(0.13)		
Student-teacher ratio	32	23.00	27	23.59	59	0.93
		(1.64)		(1.77)		
Average percentage of lessons completed	32	0.77	27	0.77	59	0.69
		(0.02)		(0.02)		
Rural (v.s. Urban)	32	0.69	27	0.67	59	0.92
		(0.08)		(0.09)		

Notes: P-values reported are from regressing the baseline characteristics on treatment (t-statistic), controlling for the strata likelihood of assignment to the interleaving condition. For the individual-level variables the standard errors are clustered by school. The school mean placement exam score is standardized by year of enrollment. Student-teacher ratio and average percentage of lessons completed are based on values from previous years. The *p-values* for female, rural, and Osun state come from z-tests. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table 2: Topics covered in end-of-term posttests and cumulative assessment

	Term 1	Term 2	Term 3	Cumulative
Panel A: Subjects covered by test				
Algebra		7%	17%	
Roman numerals			13%	3%
Geometry		23%	46%	23%
Fractions	17%	17%	7%	10%
Percentages	3%	20%	7%	10%
Measurement units		10%	7%	13%
Ratio	10%		3%	3%
Arithmetic	23%	17%		10%
Decimal bases	20%	3%		13%
HCF and LCM	27%	3%		13%
Panel B: Solution concept alignment with end-of-term posttests				
Term 1	100%			30%
Term 2		100%		37%
Term 3			100%	33%

Notes: This table reports the topics covered in each end-of-term posttest and the cumulative assessment. Material covered in the end-of-term posttests corresponds to the material covered in instruction and in-class practice during that term. Panel A provides an overview of the topics covered in each assessment. The percentage represents the percentage of questions in a specific test (e.g. posttest term 1) that covers the subtopic. Panel B shows the fraction of items on the cumulative assessment that come from each of the end-of-term tests.

Table 3: The effect of interleaved practice on follow-up and math test scores

	(1) Posttest	(2) Cumulative
Panel A: The effect of interleaved practice on follow-up		
Interleaved practice	0.006	-0.049
	(0.014)	(0.037)
	[0.701]	[0.194]
Blocked practice mean	0.96	0.78
Number of students	683	683
Number of schools	59	59
Panel B: The effect of interleaved practice on math test scores		
Interleaved practice	0.281	0.028
	(0.119)	(0.159)
	[0.022]	[0.858]
p-value (posttest effect = cumulative effect)		0.02
Blocked practice mean	0.06	0.10
Number of students	661	516
Number of schools	58	58
Panel C: The effect of interleaved practice interacted with baseline test scores		
Interleaved practice	0.348	0.021
	(0.131)	(0.205)
	[0.011]	[0.919]
Interleaved practice × Baseline test score	-0.187	-0.045
	(0.095)	(0.163)
	[0.055]	[0.782]
Baseline test score	0.699	0.552
	(0.077)	(0.120)
	[0.000]	[0.000]
Blocked practice mean	0.13	0.15
Number of students	457	362
Number of schools	43	43

Notes: Both specifications include a control for the school's linear probability of treatment given its randomization stratum. Baseline test score, in Panel C, is the average of 6 standardized baseline math scores (3 midterms and 3 endterms). Standard errors, clustered at the strata level, are reported in parentheses. Exact p-values are reported in brackets.

Appendices

Appendix A Details on the formation of interleaved practice sessions

Interleaved practice sessions varied in the amount of reviewed material and the amount of time that had elapsed between when non-aligned items had been practiced. At the start of each academic term, all practice covered the topic of the day, so there was no difference between blocked and interleaved classrooms. As the term progressed, students engaged in practicing material that was initially taught at an earlier point in time relative to the moment of practice. This means that most of the non-aligned practice was reviewed from at most a few days prior. However, by the end of the term, non-aligned material would sometimes cover material from as many as 8 units prior (a unit can range from three to five lessons), although most review was from less than 3 units prior.

NewGlobe applied a simple notation to articulate how the form of practice varied over time. The notation $L(n)$ refers to the aligned material, while $L(n-x)$ refers to material from x lessons prior. Analogously, $L(u)$ refers to material from the same unit, and $L(u-x)$ refers to material from x units prior.

In Term 1, the first seven lessons covered only the topic of the day $L(n)$. Lessons eight to ten had 50 percent $L(n)$ problems interleaved with a single problem from each of the previous 5 lessons. Lessons 11-20 had one $L(n)$ problem and nine problems each from the previous nine lessons, and the remaining thirty-two lessons used the following structure:

1. $L(n)$
2. $L(n)$
3. $L(n-1)$
4. $L(n-2)$
5. $L(n-3)$
6. $L(u-1)$

7. $L(u-2)$
8. $L(u-3)$
9. $L(n-z)$
10. Integrated problem $L(n)$ and $L(n-z)$

The penultimate question, $L(n-z)$, was selected to introduce the final, integrated problem. It was always placed at the end of the sequence of practice.

In Term 3, the first six Math 1 lessons and the first ten Math 2 lessons covered only the topic of the day $L(n)$. Starting with Lesson 11, practice typically included half $L(n)$, one item from $L(n-2)$, another from $L(n-4)$, and up to two items from prior units. In most cases, the lessons were between 1 and 3 units prior, although six practice sessions included items from four units prior, and one included items from five units prior. The practice did not include an integrated problem.

In Term 2, we do not have exact data on the interleaving strategy, although we believe that it was similar to that in Term 3.

Appendix B Proofs

Proof of Proposition 5.1

Proof. For a practice opportunity of absolute difficulty d, define

$$F_d(K) = (1-\delta)K + g(q(d,K)).$$

Using Equation 5,

$$F'_d(K) = 1 - \delta + q_K(d,K)b(q)p(q)\left[\eta_b(q) - \eta_p(q)\right].$$

The assumed bound therefore implies

$$0 \leq F'_d(K) \leq \rho < 1.$$

Consider two students with $K_t^H \geq K_t^L$ who face the same practice opportunity. By the mean value theorem,

$$K_{t+1}^H - K_{t+1}^L = F'_{d_t}(\widetilde{K}_t)\left(K_t^H - K_t^L\right)$$

for some $\widetilde{K}_t \in [K_t^L, K_t^H]$. It follows that

$$0 \leq K_{t+1}^H - K_{t+1}^L \leq \rho \left(K_t^H - K_t^L \right).$$

Iterating forward gives

$$0 \leq K_{t+s}^H - K_{t+s}^L \leq \rho^s \left(K_t^H - K_t^L \right).$$

Because $\rho < 1$, the knowledge difference converges to zero as $s \to \infty$. □

Proof of Proposition 5.2

Proof. The result follows directly from the single-peakedness of $g(q)$. If

$$q(d_0'', K_0) \leq q^*,$$

then both $q(d_0', K_0)$ and $q(d_0'', K_0)$ lie on the increasing portion of $g(q)$, and since $d_0'' > d_0'$ and $q_d(d, K) > 0$,

$$q(d_0'', K_0) > q(d_0', K_0).$$

Therefore,

$$g(q(d_0'', K_0)) > g(q(d_0', K_0)),$$

so $\Delta_1(K_0) > 0$.

If

$$q(d_0', K_0) \geq q^*,$$

then both $q(d_0', K_0)$ and $q(d_0'', K_0)$ lie on the decreasing portion of $g(q)$. Since $q(d_0'', K_0) > q(d_0', K_0)$,

$$g(q(d_0'', K_0)) < g(q(d_0', K_0)),$$

so $\Delta_1(K_0) < 0$.

Finally, if

$$q(d_0', K_0) < q^* < q(d_0'', K_0),$$

then the two regimes lie on opposite sides of the learning-maximizing difficulty. The sign of

$$g(q(d_0'', K_0)) - g(q(d_0', K_0))$$

depends on the relative heights of the learning function at the two difficulties and is therefore ambiguous. □

Proof of Proposition 5.3

Proof. Because the initial increase in difficulty is productive,

$$K_1^H - K_1^L = \Delta K_1 > 0.$$

After period 0, the two regimes face the same sequence of practice opportunities. By the mean value theorem and the assumption that $F'_d(K) \geq \lambda$,

$$K_{t+1}^H - K_{t+1}^L \geq \lambda \left(K_t^H - K_t^L \right).$$

Iterating forward gives

$$K_t^H - K_t^L \geq \lambda^{t-1} \Delta K_1$$

for every $t \geq 1$.

Since

$$\frac{\partial h(q(d,K))}{\partial K} = h'(q(d,K)) q_K(d,K),$$

the assumption

$$-h'(q(d,K)) q_K(d,K) \geq \kappa_h$$

implies that an additional unit of knowledge reduces the burden of a future practice opportunity by at least κ_h. Applying the mean value theorem again,

$$h(q_t^L) - h(q_t^H) \geq \kappa_h \left(K_t^H - K_t^L \right).$$

Using the lower bound on the knowledge difference,

$$h(q_t^L) - h(q_t^H) \geq \kappa_h \lambda^{t-1} \Delta K_1.$$

Summing over the N subsequent practice opportunities before regime L reaches mastery gives

$$\sum_{t=1}^{N} \left[h(q_t^L) - h(q_t^H) \right] \geq \kappa_h \Delta K_1 \sum_{s=0}^{N-1} \lambda^s.$$

Since $F'_d(K) \geq \lambda \geq 0$, the subsequent knowledge transition is weakly increasing in K. Hence $K_t^H \geq K_t^L$ for all $t \geq 1$, so regime H reaches the mastery threshold weakly earlier than regime L. Therefore,

$$B_h^H(\bar{K}) \leq h(q_0^H) + \sum_{t=1}^{N} h(q_t^H),$$

while

$$B_h^L(\bar{K}) = h(q_0^L) + \sum_{t=1}^{N} h(q_t^L).$$

Thus, if

$$h(q_0^H) - h(q_0^L) < \kappa_h \Delta K_1 \sum_{s=0}^{N-1} \lambda^s,$$

then

$$h(q_0^H) - h(q_0^L) < \sum_{t=1}^{N} \left[h(q_t^L) - h(q_t^H) \right],$$

which implies

$$B_h^H(\bar{K}) < B_h^L(\bar{K}).$$

□

Proof of Proposition 5.4

Proof. For any fixed number of skills m, define the total time required to master the first m skills under regime r as

$$S_m^r = \sum_{j=1}^{m} \tau_j^r.$$

Since $\tau_j^H \leq \tau_j^L$ for every affected skill j, and unaffected skills have equal mastery times across regimes,

$$S_m^H = \sum_{j=1}^{m} \tau_j^H \leq \sum_{j=1}^{m} \tau_j^L = S_m^L$$

for every m.

Let $m_L = M_L(T)$. By definition of $M_L(T)$,

$$S_{m_L}^L \leq T.$$

Since $S_{m_L}^H \leq S_{m_L}^L$, it follows that

$$S_{m_L}^H \leq T.$$

Thus, any number of skills that can be mastered under regime L within time T can also be mastered under regime H within time T. Therefore,

$$M_H(T) \geq M_L(T).$$

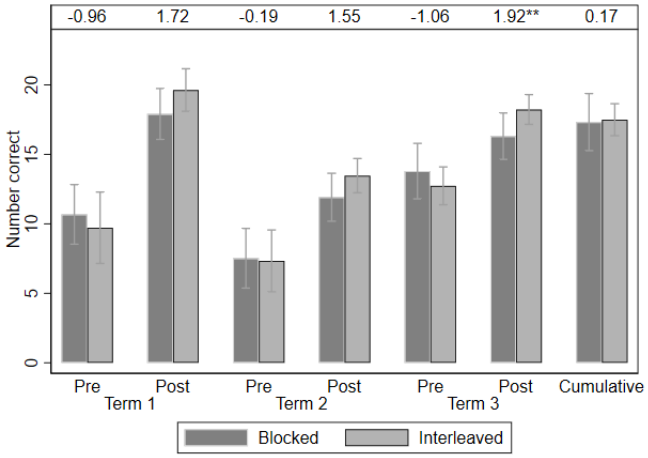
The same inequality implies the fixed-collection statement. For any fixed collection of the first m skills,

$$S_m^H \leq S_m^L,$$

so regime H requires weakly less time to master that collection. □

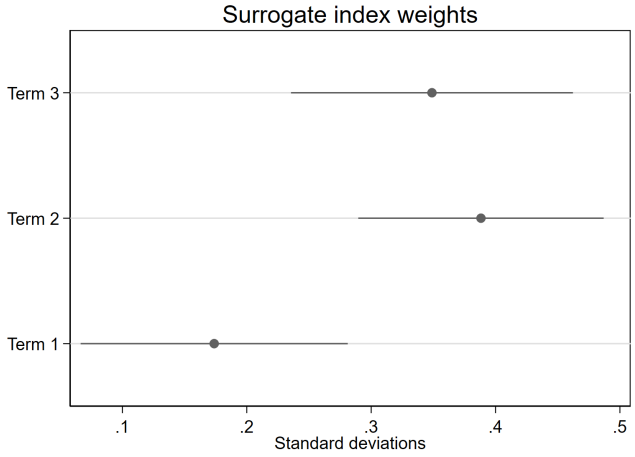
Appendix C Appendix Figures and Tables

Figure A1: Effects on number of items correct



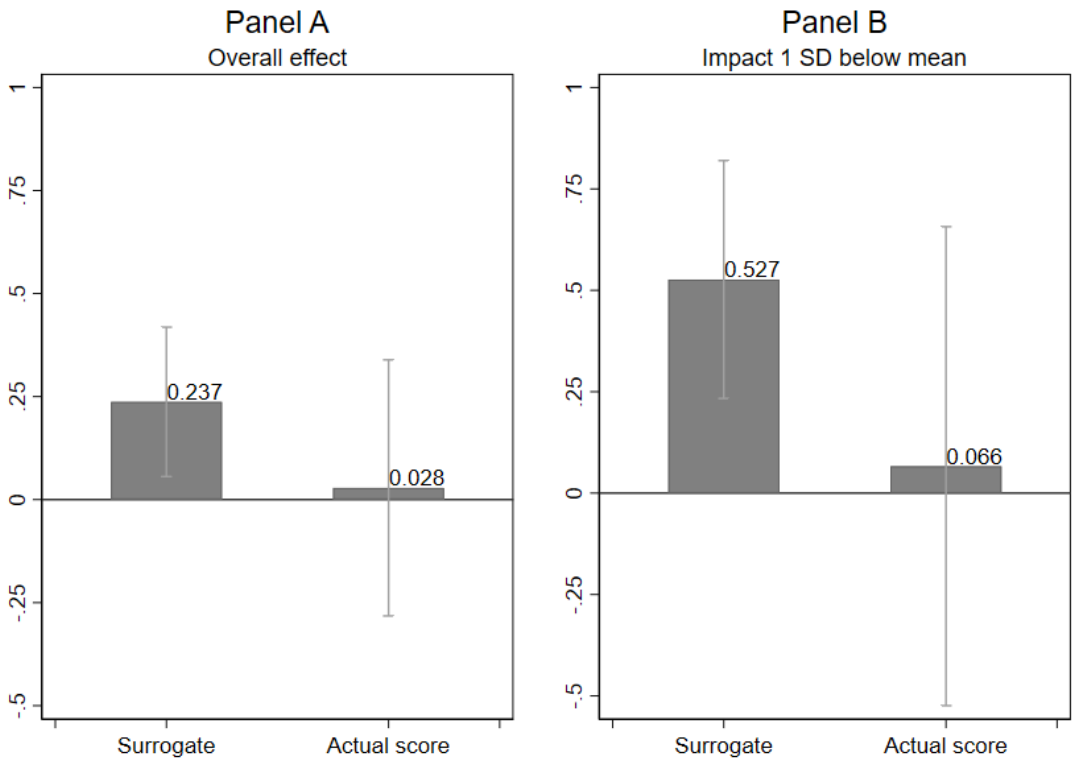
Notes: Number of items correct on each assessment. Each assessment contained 30 items. Error bars indicate 95% confidence intervals. The numbers at the top indicate the estimated interleaving effect. ***, **, and * indicate significance at 1%, 5%, and 10%.

Figure A2: Effects on number of items correct



Notes: The figure plots the coefficient estimates from estimates of Equation 3. Each coefficient estimate represents the estimated effect of a unit increase in the short-term posttest for that term on the posttest. The error bars indicate the 95% confidence interval.

Figure A3: Effect on cumulative assessment surrogate index



Notes: The figure compares what one would have expected to see on the cumulative assessment, given the impact on the termly assessments under surrogacy assumptions, to the impact observed on the actual cumulative assessment. The surrogate index is formed by estimating a linear model of the cumulative assessment as a function of performance on termly assessments in the comparison group. Standard errors for the impact on the surrogate index are obtained by bootstrapping first within schools and then within strata. Panel A reports impacts of the average treatment effect from a specification similar to Equation 1. Panel B reports impacts on initially lower performing students using a specification similar to Equation 1.

Table A1: Duration and retention interval of studies of interleaving effect for mathematical tasks

	Sample size (students) (1)	Sample size (schools) (2)	Duration of interleaving program (months) (3)	Retention interval (Days) (4)	Effect (standard deviations) (5)	Country (6)
Patel et al. (2016)	70	1	<1	<1	0.43	United States
Taylor (2008)	24	1	<1	1	1.11	United States
Taylor and Rohrer (2010b)	24	1	<1	1	1.21	United States
Patel et al. (2016)	118	1	<1	3	0.42	United States
Ziegler and Stern (2014)	72	3	<1	3	0.33	Switzerland
Ziegler and Stern (2014)	154	6	<1	3	0.46	Switzerland
Ziegler and Stern (2016)	91	5	<1	3	1.21	Switzerland
Rau et al. (2010)	54	3	<1	4	-0.58	United States
Rau et al. (2010)	54	3	<1	4	-0.41	United States
Rau et al. (2010)	54	3	<1	4	-0.34	United States
Rau et al. (2010)	54	3	<1	4	-0.56	United States
Ostrow et al. (2015)	146	1	<1	5	0.22	United States
Ziegler and Stern (2014)	72	3	<1	9	0.53	Switzerland
Ziegler and Stern (2014)	154	6	<1	9	0.50	Switzerland
Ziegler and Stern (2016)	91	5	<1	9	1.04	Switzerland
Rau et al. (2010)	54	3	<1	11	-0.83	United States
Rau et al. (2010)	54	3	<1	11	-0.12	United States
Rau et al. (2010)	54	3	<1	11	-0.69	United States
Rau et al. (2010)	54	3	<1	11	-0.28	United States
Rau et al. (2012)	115	6	<1	11	-	United States
Nemeth et al. (2021)	236	4	<1	12	0.53	Germany
Nemeth et al. (2021)	236	4	<1	19	0.41	Germany
Rohrer et al. (2014)	140	1	2	46	1.05	United States
Nemeth et al. (2021)	236	4	<1	48	0.38	Germany
Rohrer et al. (2015)	126	1	3	48	0.42	United States
Ziegler and Stern (2014)	154	6	<1	72	0.76	Switzerland
Ziegler and Stern (2016)	91	5	<1	72	0.94	Switzerland
Rohrer et al. (2015)	126	1	3	76	0.79	United States
Rohrer et al. (2020b)	787	5	4	89	0.83	United States
Ziegler and Stern (2014)	65	3	<1	92	0.38	Switzerland
Bridge Nigeria (this study)	687	59	3	36	0.28	Nigeria
Bridge Nigeria study (this study)	687	59	9	141	0.03	Nigeria

Notes: This table compares the sample size, number of schools with study participants, duration, and retention interval of studies of the impact of interleaving (versus blocked) math practice on math test scores. Duration refers to the length of time from the start of the intervention to the end (the training period). To form a comparable measure of retention intervals across studies, we standardize the calculation of the retention interval by calculating the number of days from the date of the test to the average date at which tested items were practiced as part of the intervention. Individual studies may have multiple reported impacts for separate sub-groups, alternative math assessments, and different retention intervals. Estimates from Rau et al. (2010) are the comparisons of interleaved to blocked practice reported in Brunmair and Richter (2019).

Table A2: The effect of interleaved practice on follow-up and math test scores, for stacked test scores

	(1) Posttest follow-up	(2) Posttest test scores
Interleaved practice	-0.04*	0.29**
	(0.02)	(0.12)
Blocked practice mean	0.86	0.08
Number of tests	2049	1724
Number of students	683	661
Number of schools	59	58

Notes: Specification 1 tests the effect of interleaved practice on follow-up on the stacked posttest scores, and specification 2 tests the effect of interleaved practice on stacked posttest scores. Both specifications control for the linear probability of treatment interacted with the test identifier. Standard errors are clustered at the school level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A3: Baseline balance, conditional on follow up

	Sample				
	(1) Cumulative test	(2) Posttest	(3) Posttest t1	(4) Posttest t2	(5) Posttest t3
Age	0.05	0.05	0.05	0.04	0.01
	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)
Female	0.06	0.06	0.08*	0.08*	0.08
	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)
Years enrolled	0.05	0.07	0.09	0.04	0.05
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
Baseline test	-0.08	-0.14	-0.16	-0.16	-0.10
	(0.17)	(0.17)	(0.17)	(0.18)	(0.17)
Number of students	518	665	646	582	506
Number of schools	58	58	58	57	58
Number of students with baseline test score	362	457	445	403	352
Number of schools with baseline test score	43	43	43	42	43

Notes: This table shows balance on baseline variables between interleaved and blocked conditions, conditional on follow-up on each test. Each coefficient comes from a regression of the baseline characteristic on assignment to the interleaved condition. This table also shows that one school dropped out in all samples. The reason this school dropped out is because only one student in that school meets the inclusion criteria, and this student doesn't have test scores. Standard errors are reported in parentheses. All student-level specifications control for the linear probability of treatment. Standard errors reported for student-level specifications are clustered at the school level, while those of school-level specifications are robust to heteroskedasticity. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A4: The effect of treatment on being marked present on test day

	(1) Term 1 Posttest	(2) Term 2 Posttest	(3) Term 3 Posttest	(4) Cumulative
Panel A: The effect of interleaved practice on follow-up				
Interleaved practice	0.01	-0.02	-0.06	-0.05
	(0.02)	(0.03)	(0.05)	(0.04)
Blocked practice mean	0.94	0.88	0.77	0.78
Number of students	679	682	682	683
Number of schools	59	59	59	59
Panel B: The effect of interleaved practice on having a test score				
Interleaved practice	0.00	-0.08**	-0.06	-0.05
	(0.02)	(0.04)	(0.05)	(0.04)
Blocked practice mean	0.94	0.88	0.77	0.78
Number of students	683	683	683	683
Number of schools	59	59	59	59

Notes: All specifications include a control for the linear probability of treatment. Standard errors are clustered at the school level. ***, **, and * indicate significance at 1%, 5%, and 10%. The dependent variable in Panel A is a binary variable that is equal to one if a pupil is marked as absent in the test file. The dependent variable in Panel B is a binary variable that is equal to one if the assessment was not administered by the school. The dependent variable in Panel C is a binary variable that is equal to one if a pupil is missing from a test file.

Table A5: The effect of interleaved practice on math test scores, wild bootstrap confidence intervals

	Posttest (1)	Cumulative (2)
Interleaved practice	0.28***	0.03
Wild bootstrap *p - value*	0.04	0.89
Wild bootstrap CI	[0.01, 0.61]	[-0.43, 0.43]
Number of students	661	516
Number of schools	58	58

Notes: In both specifications, we control linearly for the probability of the schools' assignment to the interleaving condition. Standard errors are clustered at the randomization strata level. P-values estimated with the wild cluster bootstrap-t procedure are provided within squared brackets below the clustered standard errors. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A6: The effect of interleaved practice on math test scores, by gender

	Posttest (1)	Cumulative (2)
Interleaved practice	0.39**	0.16
	(0.16)	(0.20)
Interleaved practice x Female	-0.20	-0.26
	(0.15)	(0.20)
Female	0.04	0.13
	(0.11)	(0.16)
Number of students	661	516
Number of schools	58	58

Notes: Female is a dummy variable that takes the value 1 if a pupil is female, and 0 if the pupil is male. In both specifications we control linearly for the probability of the schools' assignment to the interleaving condition. ATE Female reflects the average effect of interleaved practice for female students. Standard errors are clustered at the school level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A7: The effect of interleaved problem sets on test score quantiles

	(1) Posttest	(2) Cumulative
10th percentile	0.48***	0.16
	(0.14)	(0.21)
20th percentile	0.43***	0.49***
	(0.13)	(0.17)
25th percentile	0.43***	0.16
	(0.14)	(0.20)
50th percentile	0.38**	-0.00
	(0.16)	(0.19)
75th percentile	0.14	-0.09
	(0.13)	(0.18)
80th percentile	0.17	-0.16
	(0.13)	(0.17)
90th percentile	0.09	-0.33**
	(0.12)	(0.15)
Number of students	683	683
Number of schools	59	59

Notes: All specifications control for the linear probability of treatment. Each estimated effect is from a separate quantile regression. Standard errors are clustered at the school level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A8: Correlation between the different scores

	Baseline test	Posttest term 1	Posttest term 2	Posttest term 3	Cumulative test
Baseline test	1.00				
Posttest term 1	0.57	1.00			
Posttest term 2	0.46	0.48	1.00		
Posttest term 3	0.47	0.64	0.54	1.00	
Cumulative test	0.44	0.59	0.64	0.68	1.00

Notes: This table describes correlations between the different assessments within this study. The baseline test score represents the average score of six historical mid and end-term assessments administered by Bridge Nigeria in the previous year.

Table A9: The effect of interleaved practice on follow-up and math test scores
Including students who do not meet main sample inclusion criteria

	(1) Posttest	(2) Cumulative
Panel A: The effect of interleaved practice on math test scores		
Interleaved practice	0.29**	0.04
	(0.12)	(0.15)
Blocked practice mean	0.02	0.07
Number of students	729	550
Number of schools	59	59
Panel B: The effect of interleaved practice interacted with baseline test scores		
Interleaved practice	0.32**	0.01
	(0.13)	(0.20)
Interleaved practice × Baseline test score	-0.22**	-0.07
	(0.09)	(0.16)
Baseline test score	0.72***	0.58***
	(0.07)	(0.12)
Blocked practice mean	0.12	0.14
Number of students	482	379
Number of schools	45	45

Notes: This table reports analogous results to Table 3, except that it also includes test scores of students who did not meet the eligibility criteria to be included in the main analysis sample. These include students who may have enrolled after the intervention began. Both specifications include the linear probability of treatment. Baseline test score, in Panel B, is the average of 6 standardized baseline math scores (3 midterms and 3 endterms). Standard errors are clustered at the strata level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A10: The effect of interleaved practice on follow-up and math test scores
Controlling for strata fixed effects

	(1) Posttest	(2) Cumulative
Panel A: The effect of interleaved practice on follow-up		
Interleaved practice	0.00	-0.05
	(0.01)	(0.04)
Blocked practice mean	0.96	0.78
Number of students	683	683
Number of schools	59	59
Panel B: The effect of interleaved practice on math test scores		
Interleaved practice	0.29**	0.04
	(0.11)	(0.16)
p-value (posttest effect = cumulative effect)		0.02
Blocked practice mean	0.06	0.10
Number of students	661	516
Number of schools	58	58
Panel C: The effect of interleaved practice interacted with baseline test scores		
Interleaved practice	0.32**	0.03
	(0.12)	(0.20)
Interleaved practice × Baseline test score	-0.19*	0.00
	(0.11)	(0.18)
Baseline test score	0.70***	0.53***
	(0.07)	(0.13)
Blocked practice mean	0.13	0.15
Number of students	457	362
Number of schools	43	43

Notes: Both specifications include controls for randomization strata fixed effects instead of the linear control for the probability of treatment. Baseline test score, in Panel B, is the average of 6 standardized baseline math scores (3 midterms and 3 endterms). Standard errors are clustered at the strata level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A11: Lee bounds for the aggregate effect of interleaving

	Posttest (1)	Cumulative assessment (2)
Lower bound	0.278***	-0.070
	(0.105)	(0.109)
Upper bound	0.300***	0.141
	(0.114)	(0.106)
Observations	637	375

Notes: Table reports Lee bounds for the aggregate effect of interleaving on the posttest and cumulative assessments. Specification is analogous to that reported in Table 3 Standard errors are reported in parentheses. ***, **, and * indicate statistical significance at the 1, 5, and 10 percent level.