NBER WORKING PAPER SERIES

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

Working Paper 31853
http://www.nber.org/papers/w31853

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023, Revised February 2026

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

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
NBER Working Paper No. 31853
November 2023, Revised February 2026
JEL No. I20, I21, I24, I25

ABSTRACT

This paper tests whether increasing students' exposure to "desirable difficulties" improves learning in real classrooms. In a year-long field experiment in Nigerian primary schools, interleaved math practice raised short-term test performance by 0.28 standard deviations but had no effect on cumulative end-of-year assessments. Gains were concentrated among lower-achieving students, while higher-achieving students saw little or no benefit. The results suggest that strategies that make learning more effortful can boost short-term mastery but may not produce durable improvements, highlighting the limits of applying laboratory-based cognitive interventions at scale.

Michael Kremer
University of Chicago
Department of Economics
and NBER
kremermr@uchicago.edu

ⓡ

Guthrie Gray-Lobe
University of Chicago
Department of Economics
graylobe@uchicago.edu

ⓡ

Joost de Laat
University of Utrecht
j.j.delaat@uu.nl

ⓡ

Lotte van der Haar
Wageningen University & Research
lotte.vanderhaar@wur.nl

1 Introduction

A large body of research in cognitive science has documented a persistent gap between the study strategies that learners believe to be effective and those that actually promote durable learning. In laboratory conditions, the most effective approaches often involve "desirable difficulties" – conditions that make learning feel more effortful in the short run but improve long-term retention (Bjork, 1994). When students must struggle to retrieve information or apply a new concept, that effort itself strengthens memory, even when retrieval is unsuccessful (Kornell et al., 2009). In contrast, easier forms of practice that promote rapid fluency or a sense of mastery tend to produce weaker long-term outcomes.

Despite this evidence, less effortful forms of practice are commonplace (e.g. Bjork, 1994; Rohrer and Pashler, 2016; Rohrer et al., 2020a). Learners and teachers tend prefer strategies that minimize difficulty, such as blocked practice, rereading, or other more fluent study activities (Kornell and Bjork, 2008b; Karpicke et al., 2009; Yan et al., 2016). One explanation for this bias is that easy strategies appear more effective because their benefits are immediately observable: learners can quickly reproduce answers or recall facts, creating an illusion of competence. By contrast, the relationship between long-term retention and the study strategy is observed noisily, after time has passed, when the learner's memory for how they studied has faded and other study experiences may have intervened. This mismatch between short-term performance and long-term learning likely contributes to the systematic underuse of desirable difficulties in educational settings (Roediger and Pyc, 2012; Dunlosky et al., 2013a; Brown et al., 2014).

Most evidence supporting desirable-difficulty strategies, however, comes from small-scale and tightly controlled experiments. In actual classrooms, students encounter a variety of natural challenges that may also foster retention: frequent testing, cumulative review, and the need to integrate past and new material. The impacts of interleaving may also reflect behavioral responses of students, teachers or parents (Todd and Wolpin, 2003). For example, students may also self-regulate by revisiting topics they find difficult. Over the course of a school year, these ongoing and naturally occurring difficulties could help students who initially rely on less effective strategies to narrow the gap with peers who consistently engage in more effortful forms of study.

This study examines the impact of a program that increased students' exposure to desirable difficulties. Grade five classrooms in Nigerian schools were assigned to receive either blocked practice – where practice items tested the same skill (aaa-bbb-ccc) or interleaved practice, where items alternated between different skills (abc-abc-abc). We examine the impact on math test performance at the end of each academic term and over the course

of a year. The program design was informed by evidence suggesting interleaving exemplar math problems helps students better match solution strategies by learning to classify new problems (Mayfield and Chase, 2002; Taylor and Rohrer, 2010a; Rickard et al., 2008; Rohrer et al., 2014).[1]

A large literature in cognitive science has shown that interleaved practice can support the development of skills that require classification, such as applying the correct math solution to a given 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). Motivated by the strong evidence base, interleaving is often recommended as a low-cost tool to improve the quality of educational resources (Roediger and Pyc, 2012; Brown et al., 2014; Rohrer et al., 2020a; Dunlosky et al., 2013b). However, most evidence comes from short-term pilot programs that measure immediate retention. The long-term impact of sustained interleaved math practice, such as over an academic year, remains unclear. We evaluate the effect of interleaving on math test scores throughout the school year. Short-term retention was assessed with end-of-term tests covering that term's material, while long-term retention was measured with a cumulative test at the program's conclusion, covering all material taught.

Interleaved practice improved short-term retention by 0.28 standard deviations, comparable to the 0.34 standard deviation effect reported in a recent meta-analysis (Brunmair and Richter, 2019). The largest gains were among initially lower-achieving students, consistent with previous studies (e.g., Ostrow et al., 2015). Students one standard deviation below the mean score 0.54 standard deviation higher on the test when problem sets are interleaved compared to blocked.

In contrast, interleaving does not improve the retention of material over a year on a cumulative end-of-year-assessment. The average effect on this test, which covers the same material as short-term assessments, is just 0.03 standard deviations and indistinguishable from zero. This finding is remarkable given the robust gains from interleaving observed on posttests. We apply surrogate analysis (Prentice, 1989; Athey et al., Forthcoming) to show that the impact one would expect to find from interleaving, given the gains observed at the end of each term is 0.24 standard deviations. These results suggest that the literature on interleaving (and perhaps other desirable difficulties) may overstate the gains we would expect over the course of a year-long increase in the use of desirable difficulties.

Quantile regression estimates suggest that interleaving may benefit lower-achieving students more persistently, but these gains are offset by negative effects for higher achievers. These findings suggest that the order in which math skills are practiced may have distribu-

[1]The design of the program was specifically informed by Rohrer (2017), an evidence-based practitioner guide.

tional consequences and that the effective use of interleaving may require targeting practice to individual student need, similar to the way many interventions aim to target the level of instruction to student performance (Banerjee et al., 2007; Raudenbush et al., 2020; Angrist and Meager, 2023).

This study is the first to experimentally evaluate an interleaving program implemented over a full academic year. It builds on prior cognitive science research examining how interleaved practice influences long-term retention of mathematical skills but differs in scope and intent: rather than a tightly controlled laboratory intervention, it represents a practitioner-led implementation integrated into regular instruction over an academic year. The assessments capture effects across a wide range of retention intervals – from material introduced more than 280 days before testing to material taught only 12 days prior – with an average interval of roughly 145 days on the cumulative assessment. Previously, the longest interleaving program experimentally tested in mathematics lasted four months in 52 Florida schools and measured retention after one month, finding an effect of 0.83 standard deviations (Rohrer et al., 2020b), while a two-day instructional intervention evaluated ten weeks later produced an effect of 0.34 standard deviations (Ziegler and Stern, 2014).

This study adds to the emerging literature on using cognitive psychology to design scalable, low-cost educational interventions in low- and middle-income countries (LMICs) (Dillon et al., 2017). Evidence of interleaving's impact on math skills comes mainly from researcher-led pilot studies in the United States, Germany, or Switzerland (see Appendix Table A1). These findings may not generalize to other contexts due to differences in populations, baseline learning levels, and complementary inputs. Impacts from researcher-driven pilots may also not translate to practitioner-led programs at scale (Al-Ubaydli et al., 2019; List, 2022). This study shows that practitioners can replicate pilots' short-term gains, but raises questions about the long-term persistence of these effects.

While short-term impacts may overstate long-term gains, interleaving can still be a valuable tool. First, the present study cannot rule out moderate positive aggregate effects on long-term retention, and manipulating practice sequences is often low-cost. Second, substantial short-term test score gains may have intrinsic value. Although our study fixed the pace of instruction, if interleaved practice supports faster learning – especially among lower-performing students – interleaving could enable an accelerated pace of instruction. Third, heterogeneous effects suggest that more effective practice might be developed by targeting based on students' skills. A positive policy implication of the study is that to effectively exploit lessons from the study of memory may require complementary changes to education, such as accelerating the pace of instruction or targeting study strategies based on their expected gains. While the need for more comprehensive change to exploit desirable difficulties

in the classroom may erode the cost-effectiveness argument for such strategies, such interventions may still be feasible at low cost in some settings. We view this as an important area of future research. The present study is meant to examine whether increased exposure to desirable difficulties over the year can improve math outcomes on its own.

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 empirical framework we use to test the effect of the interleaved practice strategy on math performance. Section 4 presents the results. Section 5 interprets the results and concludes.

2 Context

This section describes the context and details of the blocked and interleaved practice programs. Section 2.1 covers the background, setting, math curriculum, and in-class practice sessions. Section 2.2 explains how the interleaving program modified practice sessions. Subsection 2.4 describes the randomization procedure and inclusion criteria. Sub-section 2.5 describes the study population. Sub-section 2.6 discusses the test scores used to evaluate the program.

2.1 Background & setting

The interleaved versus blocked practice program was implemented in grade five math 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 month per student.

The interleaving experiment was initiated by Bridge to address student difficulties in classifying math problems (matching appropriate solution concepts to problems) on cumulative assessments.[2] Bridge staff responsible for designing lesson materials consulted cognitive science research to identify methods for improving retention, particularly in categorizing math problems and applying correct solution strategies.

[2]Bridge schools administer up to seven subject-specific tests per year, including an initial diagnostic, midterm, and endterm assessments for each of the three academic terms. These assessments provide data for centrally monitoring lesson material performance.

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 math instruction daily. New material is introduced daily; both periods typically cover similar content and may involve different types of problems.[3] 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 math period includes two "my-turn/your-turn" sessions, covering the same topic with different problems.

Students in Bridge schools have numerous opportunities to review material. Weekly mathematics revision courses review material from the past week. Students are tested at mid- and end-term, providing additional opportunities for retrieval practice (Karpicke and Roediger, 2008). Within each term, the midterm assessment took place before the posttest, and the endterm took place after the endterm. The term 3 endterm also took place after the cumulative assessment. Both midterms and endterms contain a mix of material from the present and past terms.[4]

2.2 Design of the interleaving program

The interleaving program modified the sequence of problems presented to children during the daily "my turn, your turn" module in both Mathematics periods.[5] The number and type of practice problems completed by the teacher during "my turn" were identical in both

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

[4]This study does not report estimates of the effect on midterm and endterm assessments because they contain items taught to students before the intervention. This study focuses on the assessments that were administered to specifically measure skills for which there is experimental variation in the form of practice. We find some evidence of gains on the midterm and endterm assessments, especially in Term 2, and not in Term 3.

[5]In Term 1, only Mathematics 1 practice was blocked in both groups. In Terms 2 and 3, interleaved classrooms received interleaved practice in both daily Mathematics 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.[6] Figure 1 provides examples of problems used in blocked and interleaved classrooms.

In the interleaved condition, most 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 or endline assessments tend to encompass a more diverse set of topics than midline tests.

The interleaving program was not designed to isolate the effect of interleaving *per se*. Because the intervention altered multiple aspects of practice sequencing, effects cannot be attributed solely to interleaving. Interleaving manipulates both the spacing (the amount of time between practice attempts) and degree to which students must discriminate between problems (Kornell and Bjork, 2008a; Taylor and Rohrer, 2010a). Laboratory studies have attempted to disentangle interleaving from related features such as spacing or retrieval practice, but such designs are rarely feasible in real-world classroom settings (Taylor and Rohrer, 2010a). 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). While we believe this grade 3 reading program likely did not affect grade 5 outcomes, this cannot be empirically tested.

[6]Details in Appendix A.

2.3 Overall research agenda

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 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).

2.4 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. Entry and exit of students 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

[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.

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 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).

2.5 Sample characteristics & baseline balance

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 in rural areas, though many Lagos schools might be better described as peri-urban. Eight percent are in Osun state, a less urban area, where all schools are classified as rural.

2.6 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

[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.

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. Afterward, students received the correct answers. Teachers transmit the count correct 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. Items were mirror versions of problems drawn uniformly from the "your turn/my turn" practice sessions throughout the year. 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

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

[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, increases follow-up rates (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.

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 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 Empirical framework

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

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

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 whether the classroom was assigned to the interleaving group, p_j is the probability that school j would be assigned to the interleaving condition, and ϵ_{ijs} is an idiosyncratic error term potentially containing a common classroom component.[16] 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.[17]

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 variation in the impact of interleaving on students is predicted by a student's baseline test score 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 group. We use the mean of the three midterm and endterm math scores before the start

[16] We control for the probability that the school would have been assigned to the interleaving condition instead of strata dummies because, for specifications in this study, there are randomization strata with non-varying treatment status after conditioning on observation of all data used in the specification. This is especially important in the case of the tests of heterogeneous impacts across students with different baseline test scores because these data are not available from all schools. Controlling for the probability of treatment is sufficient to ensure unconfoundedness and conserves the sample size for estimation. Unconfoundedness follows from Theorem 1 of (Rosenbaum and Rubin, 1983). To see this note that $Y(1), Y(0) \perp Z|X$, where X represents randomization strata fixed effects is true given randomization, and therefore $Y(1), Y(0) \perp Z_i|e(X)$, where $e(X)$ is the probability of treatment given a units randomization stratum. Results are broadly similar when controlling for strata fixed effects (see Appendix Table A10).

[17] 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 2 (out of 10) strata contain more than 10 schools (Appendix Figure A1). 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).

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 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 the effect *on* students at different initial rankings within the classroom. 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.

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 (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 others 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 (Table 3, Panel B, Column 2). The null hypothesis that the interleaving effect on the cumulative assessment is zero cannot be rejected. Given the standard error, moderate positive or negative impacts of interleaving on the cumulative assessment cannot be ruled out. The 95 percent confidence interval includes effects from -0.28 to 0.34 standard deviations.

The longer-term interleaving effect is smaller than the short-term effect (Table 3, Panel B)[18]. A test of the hypothesis that the effects on the posttest and cumulative assessment are equal yields a p-value of 0.02.

4.2 Surrogate analysis

What long-run effect would we expect on the cumulative assessment, given the observed short-term impacts on termly posttests? We explore this question applying procedures from surrogate analysis (Prentice, 1989; Athey et al., Forthcoming).

Surrogate analysis is designed to forecast treatment effects on long-term outcomes when (a) treatment effects are observed on short-term (surrogate) measures, and (b) the relationship between short- and long-term outcomes can be 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 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 study, both short- and long-term effects have been observed, so the value of short-term effects as surrogates can be evaluated. The results show that short-term test scores provide poor predictions of the cumulative assessment and are therefore weak surrogates.

As would be expected, termly posttests are highly predictive of the cumulative assessment. 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

[18]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.

group

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

We then estimate Equation 1 using the prediction from Equation 3, $\hat{Y}_{ij,\text{cumulative}}$ (the surrogate index) as an outcome. Under the assumptions needed for OLS to identify the causal effect of each termly posttest (Chapter 4 Wooldridge, 2010), the effect of interleaving on the surrogate index reflects the impacts 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 surrogate index.

Estimates of β_t from Equation 3 are reported in Appendix Figure A2. Consistent with the view that the posttests capture a distinct set of skills, the estimates show that each termly score is separately predictive of performance on the cumulative assessment.

In aggregate, the estimated 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 7). The gap is especially striking when considering initially lower-performing students. Panel B of Figure 7 also reports the impacts using both the surrogate 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 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. However, in light of the

heterogeneity on the posttest, it is noteworthy that these results are similarly signed to those for the posttest.

The effect 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.[19] The estimated effect on the 20th percentile is 0.49 standard deviations and highly statistically significant (Figure 6) on the long-term test.[20]. Notably, 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.

If one does not accept the rank preservation assumption discussed in Section 3, then the negative impacts on students who would have been at the top of the distribution in the counterfactual would be even larger.

4.4 Isolating impacts on long-term retention

The absence of an effect on the cumulative assessment, coupled with the large positive impacts observed in the final term, suggests small or even negative impacts on long-run retention. Denote a student's latent skill related to material from term t by θ_t at the time of the cumulative assessment. Approximately one-third of the items on the cumulative assessment mirrored those in the posttest, which directly measured θ_3.

To illustrate, consider a potential outcomes model of the latent skills that influence performance on the cumulative assessment. Assume that the cumulative assessment score is given by

$$y_i = \sum_{t=1}^{3} \omega_t \theta_{ti}(0) + Z_i \sum_{t=1}^{3} \left[\omega_t \underbrace{(\theta_{ti}(1) - \theta_{ti}(0))}_{\delta_{ti}} \right]$$

where ω_t represents the fraction of items that require skills from the term t domain, and $\theta_{t,i}(z)$ represents individual i's latent skill as a potential outcome of z (Rubin, 1974). The difference in potential latent skills $\theta_{ti}(1) - \theta_{ti}(0)$, represents the impact of interleaving on test items related to term t material. The average cumulative assessment interleaving effect is $\rho = E[\sum_t \omega_t \delta_t]$. In other words, if the effect is 0.3 on θ_3, and the cumulative effect is 0.03, then the impact on retention of term 1 and 2 material is $\rho - 0.33\delta_3$, or -0.06.

This same exercise suggests even larger and more statistically significant negative impacts on higher-performing students. Figure 6 shows two approaches to eliminate the influence on the Term 3 posttest from the cumulative assessment. The bottom left panel reports

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

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

quantile regression estimates using a residual performance on the cumulative assessment after removing the number we expect they would have correctly answered, given their performance on the Term 3 posttest. Because 33 percent of the cumulative assessment included items that would have been overlapping with the Term 3 assessment, we form a residualized test score: $\tilde{y} = \text{Cumulative score} - 0.33 \times \text{Term 3 posttest score}$. The bottom right panel shows results from the quantile regression on the OLS residual cumulative score after controlling for term 3 posttest scores. In both cases, estimates at all quantiles are shifted downwards, and estimated effects on the top quartile are large and in many cases highly statistically significant.

The assumptions needed for these analyses to yield an unbiased estimate of the causal effect on skills related to terms 1 and 2 material are strong. First, it must be the case that the term 3 posttest represents skills on term 3 material at the time of the assessment. If the posttest, which provided distributed and in some cases interleaved practice of term 3 material, had a pedagogical effect, it is possible that despite being close in time, students in the comparison group improved as a result of the posttest. In this case, controls for posttest performance will lead to negative bias in the estimates of the impact on term 1 and 2 material. Second, it must be the case that impacts on common skills across the terms are minimal. If, for example, interleaving produced gains on transferable conceptual skills, then the term 3 posttest would be a "bad control" (Angrist and Pischke, 2008).

4.5 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.

4.5.1 Catch up or 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,

making it unlikely that consolidated skills would be quickly lost. Additionally, students in blocked conditions performed better on the cumulative assessment than on the Term 2 or Term 3 assessments (Appendix Figure A3), further suggesting retention rather than decay.

4.5.2 Unit inconsistency

The cumulative assessment may reflect a comparison of inconsistent units. A standard deviation of the cumulative assessment distribution may measure a larger difference in learning than a standard deviation of the posttest distribution. This might be because the test is more comprehensive or due to differences in the tests' properties. 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. This interpretation does not have strong support in the data: the cumulative assessment is similarly correlated with posttests as the other posttests are to one another (See Appendix Table A8).

4.5.3 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 impacts on the Term 3 posttest (Figure 5).

4.6 Gender heterogeneity

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

5 Conclusion

This study evaluates a full academic-year program that induced "desirable difficulties" in grade 5 math classes in Nigeria. In the interleaved program, students practiced problems drawn from different days' lessons, creating a mix of content that is generally considered more challenging. Students in comparison classes practiced only problems aligned with that

day's specific lesson. Interleaving improved short-term retention on term-specific tests, with effects similar to those found in high-income countries. However, interleaving did not improve long-term retention as measured by performance on the end-of-year cumulative test.

These results can be reconciled with the existing literature if one recognizes the opportunities students in the comparison group may have had to catch up with those who got a head start from interleaved practice. Scientific studies of memory typically control counterfactual learning conditions to establish clear causal connections between different learning processes. However, these conditions may overlook that many desirable difficulties – such as distributed or interleaved practice – can arise naturally in educational settings, and that students may also adjust their study effort to compensate for less effective practice.

First, Bridge schools, like many other education systems, provide students with many opportunities to be tested. Including the termly posttests used to evaluate the impact of the interleaving program, students in this study were tested up to eight times before taking the cumulative assessment. Each of these tests provided students with opportunities for distributed practice, another validated "desirable difficulty" (e.g. Ebbinghaus, 1913; Bjork, 1994; Brown et al., 2014). As is standard practice for mid and endterm assessments, these tests themselves would have included many types of items, potentially capturing some of the benefit of interleaving. Insofar as the memory-formation benefit of such practice would be greater for those for whom the practice was more difficult (Pavlik and Anderson, 2005), it is possible the testing apparatus would have larger impacts on the retention of those who initially formed less durable memories using blocked practice.

Second, over time, students may seek out additional opportunities to learn things that they don't initially master. A canonical result of education economics due to Todd and Wolpin (2003) is that the impact of an intervention reflects the combined effect on behaviors of students, parents, and teachers. While responses may vary, students are not passive recipients of information. They make decisions about how to pay attention, how much to study before a test, and which subjects to focus on. Teachers and parents may also respond by trying to remediate when students do not learn skills initially. Endogenous effort responses could also explain why interleaving is less effective for higher-performing students, as these students may be those for whom this endogenous response is greatest (e.g., self-directed learners).

The absence of an aggregate impact on the long-term retention test may also reflect heterogeneous distributional impacts of interleaving. Interleaving had the largest impact on termly assessments among students at the bottom of the distribution. Lower-performing students may have benefited in the long-term as well. However, insofar as there is evidence of gains at the bottom, there is evidence of harm at the top of the class. These results, while

not dispositive, suggest that efforts to match students to practice that is more suited to their needs may be valuable. A large literature has emerged on the potential benefits of targeting instruction to the level of individual students (e.g. Banerjee et al., 2007; Duflo et al., 2011; Banerjee et al., 2017; Muralidharan et al., 2019; Raudenbush et al., 2020). Typically, this involves assigning more difficult work to students who appear more prepared. In this case, it appears that lower-performing students benefit the most from more challenging forms of practice.

This study is unable to rule out some aggregate long-term gains from interleaving. The 95 percent confidence interval of the aggregate effect of interleaving contains both moderate positive and negative effects. Overall, the study is inconclusive regarding the question of whether interleaving or blocked practice is a better form of math practice when it comes to cumulative math performance. The gains from interleaving in the short term suggest that, at a minimum, interleaving produces more rapid learning over some time horizons. For example, if students who struggle with math early are more likely to develop math anxiety (Ramirez et al., 2013) or a fixed mindset (Blackwell et al., 2007), then intervention to accelerate their progress may have long-term impacts on their self-perception, relationship with math, and effort toward math mastery. The present study is, unfortunately, not positioned to examine such a hypothesis.

The large short-term impacts on the bottom of the distribution suggest that combining interleaving with other changes to the classroom environment may be very effective. The pace of group-based math instruction can be slowed when lower-performing students are struggling to keep up. These results suggest that interleaving may reduce the prevalence of students falling behind and thereby allow for accelerated instruction, including more advanced material that may benefit higher-achieving students (Cohodes, 2020). There may also be unobserved benefits in the present study of helping lower-performing students learn more initially.

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.

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, J. D. and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Angrist, N. and R. Meager (2023): "Implementation matters: Generalizing treatment effects in education," *SSRN No. 4487496.* Available at SSRN: `https://dx.doi.org/10.2139/ssrn.4487496`.

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

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017): "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, 31, 73–102.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007): "Remedying Education: Evidence from Two Randomized Experiments in India," *The Quarterly Journal of Economics*, 122, 1235–1264.

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.

Blackwell, L. S., K. H. Trzesniewski, and C. S. Dweck (2007): "Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention," *Child Development*, 78, 246–263.

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.

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.

Cohodes, S. R. (2020): "The Long-Run Impacts of Specialized Programming for High-Achieving Students," *American Economic Journal: Economic Policy*, 12, 127–66.

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.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101, 1739–74.

Dunlosky, J., K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham (2013a): "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.

——— (2013b): "Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology," *Psychological Science in the Public Interest*, 14, 4–58.

Ebbinghaus, H. (1913): *Memory: a Contribution to Experimental Psychology*, New York, NY: Teachers college, Columbia University.

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.

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., A. C. Butler, and H. L. R. III (2009): "Metacognitive strategies in student learning: Do students practise retrieval when they study on their own?" *Memory*, 17, 471–479.

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

——— (2008b): "Learning concepts and categories: Is spacing the "enemy of induction"?" *Psychological science*, 19, 585–592.

Kornell, N., M. J. Hays, and R. A. Bjork (2009): "Unsuccessful retrieval attempts enhance subsequent learning," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 989–998.

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.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019): "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 109, 1426–60.

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*.

Pavlik, P. I. J. and J. R. Anderson (2005): "Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect," *Cognitive Science*, 29, 559–586.

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

Ramirez, G., E. A. Gunderson, S. C. Levine, and S. L. Beilock (2013): "Math Anxiety, Working Memory, and Math Achievement in Early Elementary School," *Journal of Cognition and Development*, 14, 187–202.

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.

Raudenbush, S. W., M. Hernandez, S. Goldin-Meadow, C. Carrazza, A. Foley, D. Leslie, J. E. Sorkin, and S. C. Levine (2020): "Longitudinally adaptive assessment and instruction increase numerical skills of preschool children," *Proceedings of the National Academy of Sciences*, 117, 27945–27953.

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.

Rohrer, D. and H. Pashler (2016): "Recent Research on Human Learning Challenges Conventional Instructional Strategies," *Educational Researcher*, 39, 406–412.

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.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688–701.

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.

Todd, P. E. and K. I. Wolpin (2003): "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113, F3–F33.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2nd ed.

Yan, V. X., E. L. Bjork, and R. A. Bjork (2016): "On the Difficulty of Mending Metacognitive Illusions: A Priori Theories, Fluency Effects, and Misattributions of the Interleaving Benefit," *Journal of Experimental Psychology: General*, online first publication, May 26, 2016.

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.

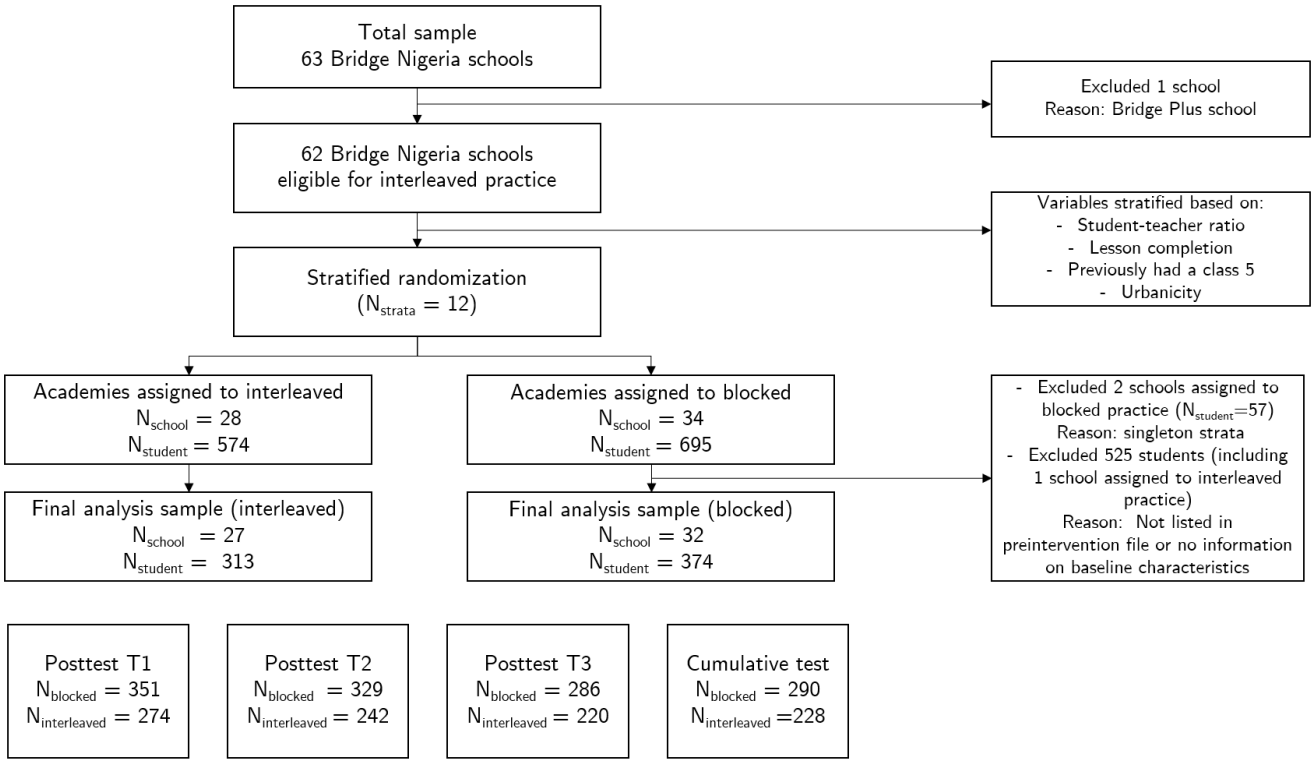
Figures & Tables

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?

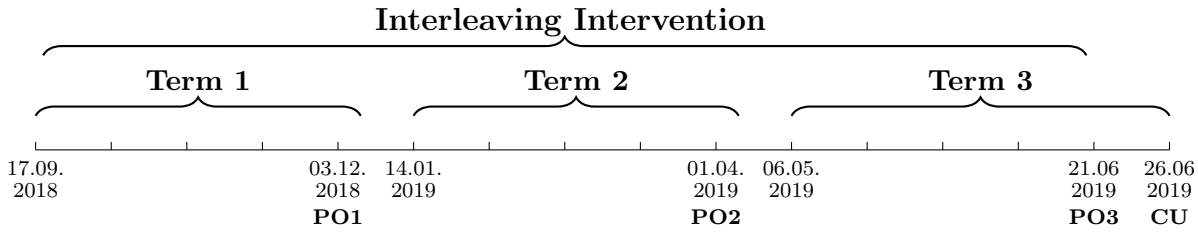
Notes: Examples of blocked practice vs. interleaved practice (independent study). All items in the blocked practice set cover the topic of the day. Items in the interleaved practice set cover the topic of the day (problem 1, 3, 5, 8), topics from previous lessons (problem 2, 4, 6) and an integrated problem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

Figure 2: Consort diagram



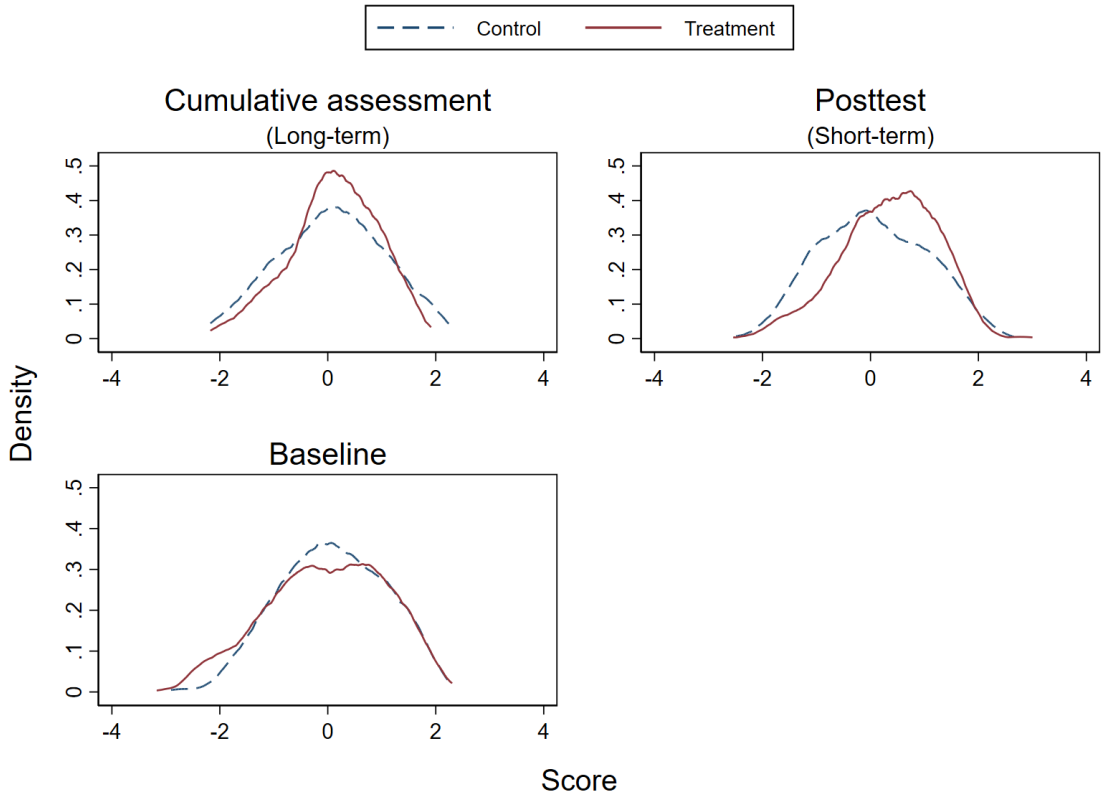
Notes: Figure shows the process of arriving at the final sample of schools and students for analysis. The final row shows the number of students in blocked and interleaved schools that have a test for each of the tests.

Figure 3: Timeline of events



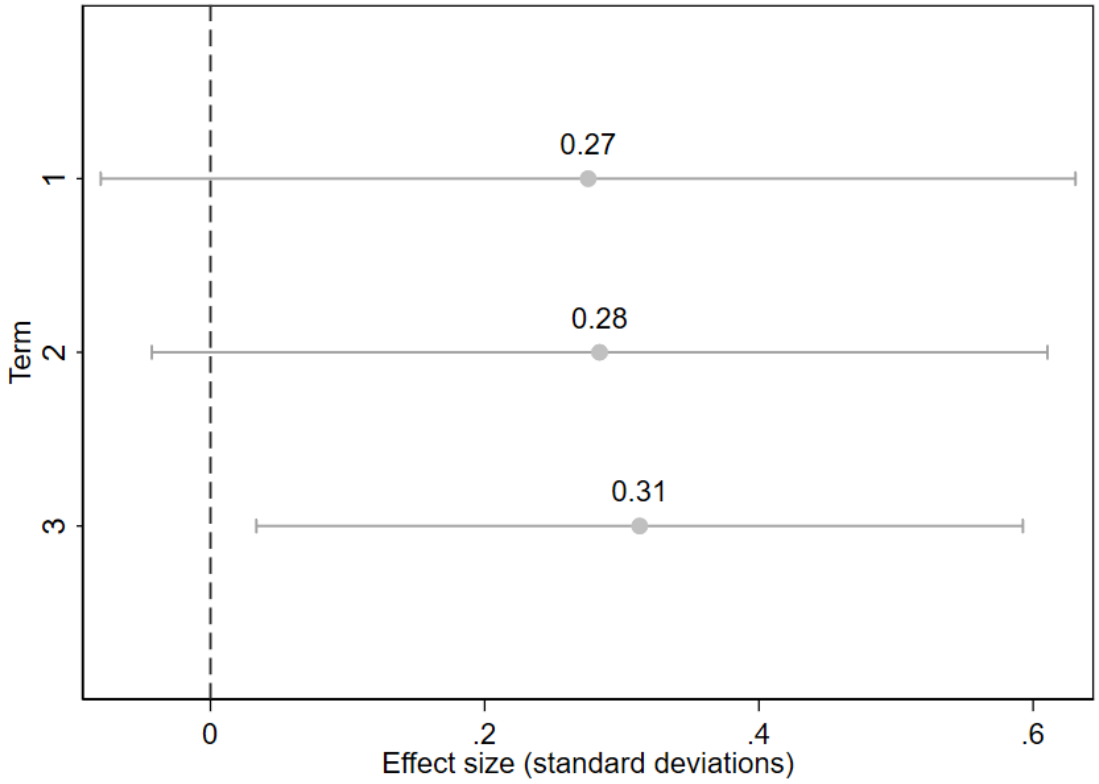
Notes: Figure shows the project timeline. PO=Posttest and CU=Cumulative assessment.

Figure 4: Test score distributions



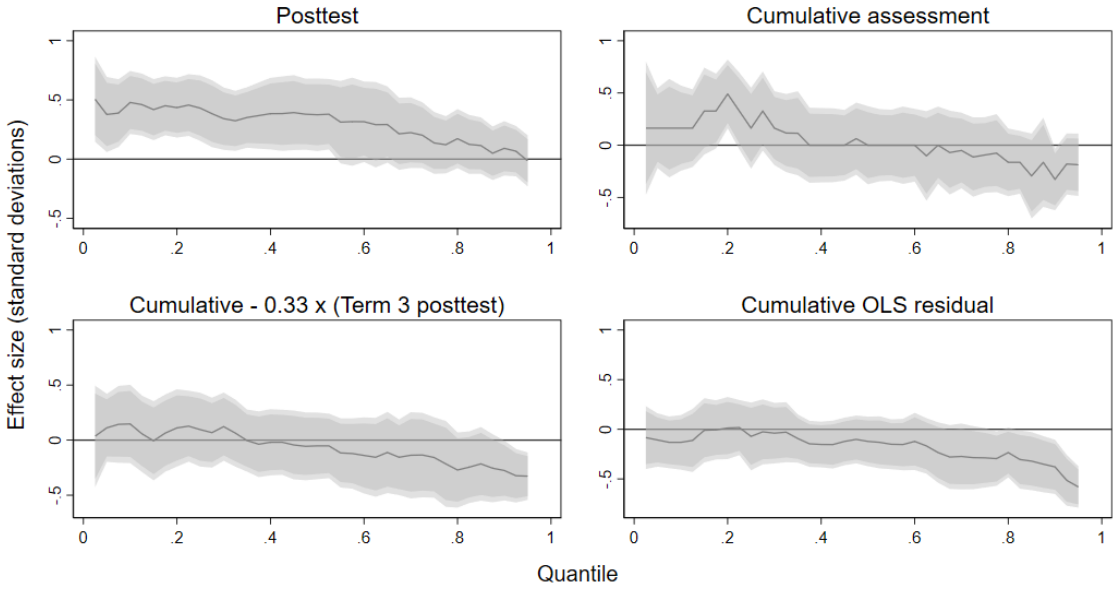
Notes: Each figure plots the kernel density of test scores in blocked and interleaved classrooms.

Figure 5: Effect on individual posttests



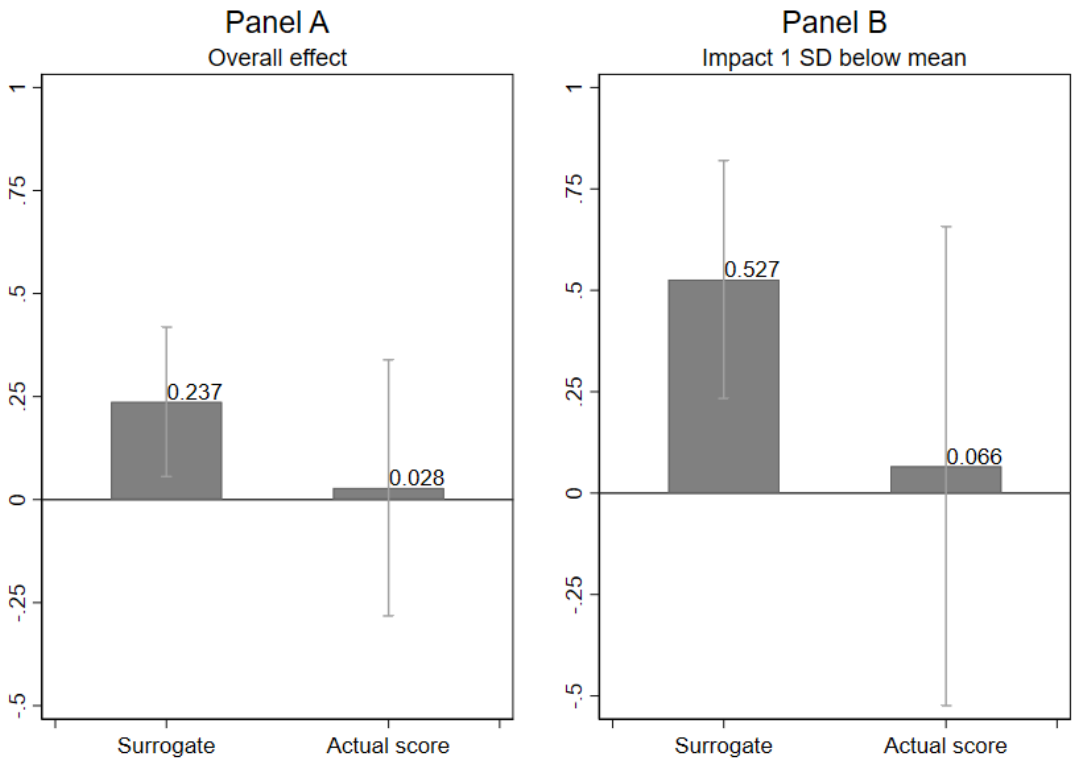
Notes: Coefficient plot of impacts on individual posttests. Error bars indicate 95% confidence interval.

Figure 6: Effects across quantiles



Notes: The figure shows the effects of interleaving at percentiles of the distribution of test scores. 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: 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 1: Table 2: 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	0.01	-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	0.28**	0.03
	(0.12)	(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	0.35**	0.02
	(0.13)	(0.21)
Interleaved practice × Baseline test score	-0.19*	-0.05
	(0.09)	(0.16)
Baseline test score	0.70***	0.55***
	(0.08)	(0.12)
Blocked practice mean	0.13	0.15
Number of students	457	362
Number of schools	43	43

Notes: 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%.

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 review 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 unitFs prior (a unit can range from between 3 and 5 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 1 $L(n)$ problem and 9 problems each from the previous 9 lessons and the remaining 32 lessons included 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 $L(n-z)$

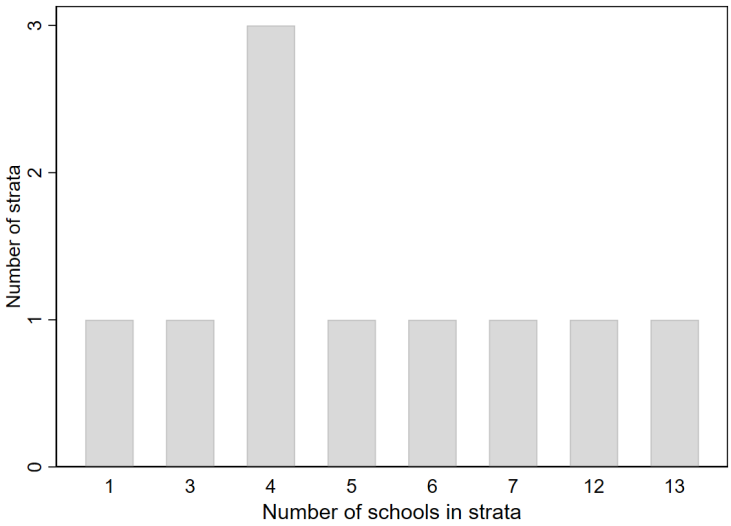
The penultimate question $L(n-z)$ was selected to introduce the final, integrated problem. It always was placed at the end of the sequence of practice.

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.

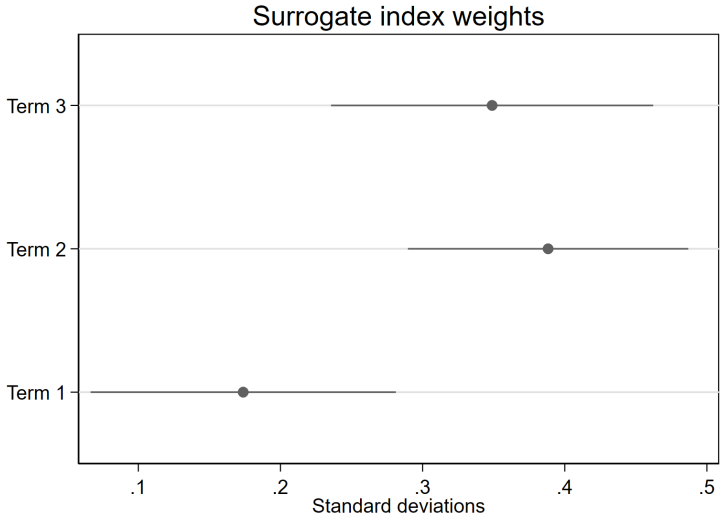
Appendix B Appendix Figures and Tables

Figure A1: Number of schools per strata



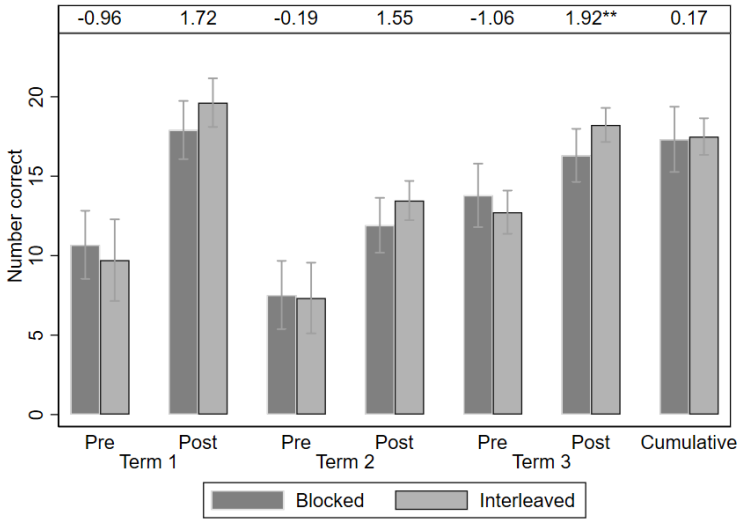
Notes: This figure graphs the number of schools (units of randomization) per randomization stratum. The horizontal represents the number of schools in the stratum, and the vertical axis represents the number of strata with the corresponding number of schools.

Figure A2: Effects on number of items correct



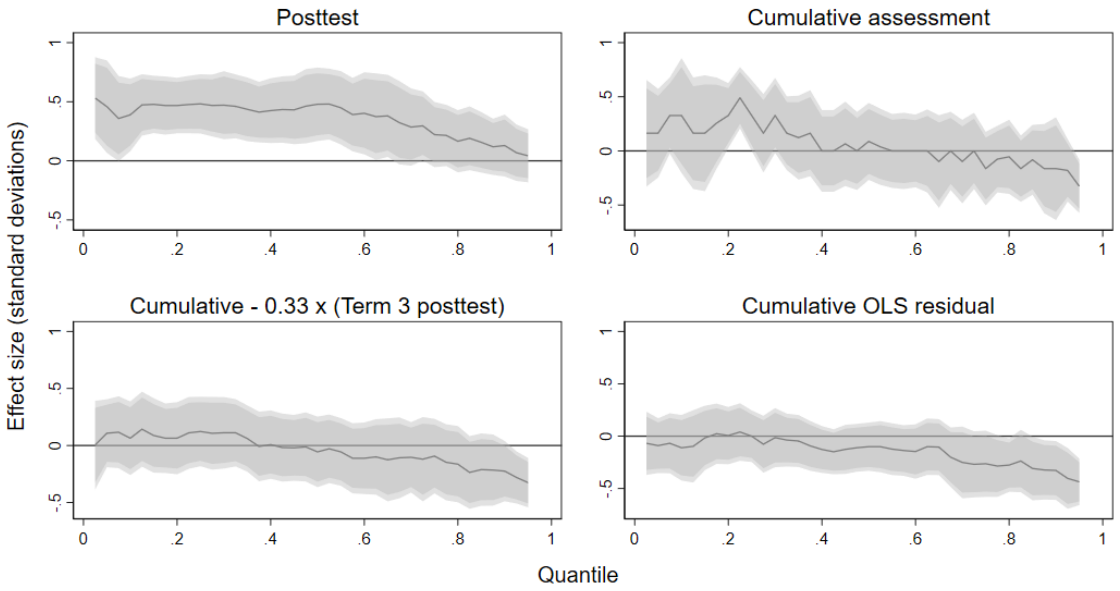
Notes: The figure plots the coefficient estimates from estimates of Equation 3. Each coefficient estimate represents the estimated effect of a unit increase in the short-term posttest for that term on the posttest. The error bars indicate the 95% confidence interval.

Figure A3: 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 A4: Effects across quantiles



Notes: The figure presents results analogous to those in Figure 6, but including those students who do not meet the inclusion criteria for the main analysis sample. The Figure shows the effects of interleaving at percentiles of the distribution of test scores. 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.

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)	(0.108)
Upper bound	0.300***	0.139
	(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.