NBER WORKING PAPER SERIES

THE LONG-TERM DISTRIBUTIONAL IMPACTS OF A FULL-YEAR INTERLEAVING MATH PROGRAM IN NIGERIA

Lotte van der Haar
Guthrie Gray-Lobe
Michael Kremer
Joost de Laat

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

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

The evaluation received support from the J-PAL Post-Primary Education Initiative and the International Growth Centre (IGC). This study reports results from an experiment designed and conducted in schools operated by NewGlobe. We thank Shannon May, Sean Geraghty, Tim Sullivan, Daniel Rodriguez-Segura, and Clotilde de Maricourt for the collaboration and providing valuable feedback. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Long-Term Distributional Impacts of a Full-Year Interleaving Math Program in Nigeria
Lotte van der Haar, Guthrie Gray-Lobe, Michael Kremer, and Joost de Laat
NBER Working Paper No. 31853
November 2023
JEL No. I20,I21,I24,I25

ABSTRACT

This study reports the findings from a year-long randomized evaluation assessing the impact of assigning 62 classrooms in Nigeria to receive either blocked or interleaved math problem sets. Blocked practice sessions focused on a single skill at a time. Interleaved problem sets alternated between different skills within a practice session. On tests of short-term retention, interleaved practice increased test scores by 0.29 standard deviations. In contrast, we find no evidence that interleaving improves average performance on a cumulative assessment measuring retention of material over the academic year. We find some evidence of large impacts on the cumulative assessment at the bottom of the distribution, but these impacts appear to be offset by negative impacts at the top.

Lotte van der Haar
Wageningen University
Development Economics Group
lotte.vanderhaar@wur.nl

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

Michael Kremer
University of Chicago
Department of Economics
1126 E. 59th St.
Chicago, IL 60637
and NBER
kremermr@uchicago.edu

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

1 Introduction

A large literature in cognitive science has shown that interleaved practice – where the sequence of material alternates between topics (abc-abc-abc) – can support inductive learning, but there is little evidence on the impact of educational policies that increase exposure to this mnemonic tool. Interleaving contrasts with blocked problem sets, where students repeatedly practice a single skill (aaa-bbb-ccc). Interleaving has been found to support the development of skills requiring classification, such as applying the correct math solution to a given problem (e.g., Taylor and Rohrer, 2010; 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). Most evidence is based on high-quality pilot programs that increase exposure to interleaving over relatively short periods and measure retention over shorter terms. The impact of programs that increase exposure to interleaved math practice over long terms, such as an academic year, remains uncertain.

This study examines the effects of interleaved math problem sets relative to blocked problem sets on math test scores in low-cost private schools in Nigeria, operated by Bridge Nigeria. Grade five classrooms were randomly assigned to practice math concepts following either an interleaved or blocked sequence for a full academic year. The design of the program was guided by a large body of evidence suggesting that interleaving exemplar math problems could help students learn to more accurately match solution strategies to problems through an improved ability to accurately classify new problems and improve long-term retention of these skills.

We examine the effect of interleaving on multiple mathematics test scores observed throughout the program. Short-term retention was measured using assessments at the end of each academic term that tested material taught in that term. Longer-term retention was measured using a cumulative assessment at the end of the program that covered material taught over the program's full course.

Interleaved practice increased short-term retention by 0.29 standard deviations. This effect is close to the 0.34 standard deviation effect of interleaving on mathematics skills found in a recent meta-analysis by Brunmair and Richter (2019). The interleaving effect is largest for children at the bottom of the test score distribution, also similar to prior studies of interleaved practice (e.g., Ostrow et al., 2015). We do not find that interleaving improved test scores of top-performing students.

In contrast, we find no evidence of an average effect on a measure of retention of material over the full academic year. The estimated effect of interleaving on an end-of-year cumulative assessment is only 0.04

standard deviations. We cannot reject the hypothesis of zero effect. Quantile regression estimates suggest that the interleaving effect may be more persistent for lower-achieving students, however, these impacts appear to be offset by negative impacts at the top of the distribution. These results suggest that the order of practice can have distributional consequences.

Distributional impacts of the order of practice may reflect differences in students' counterfactual ability to accurately classify math problems. Low-performing students may struggle to accurately classify problems, a form of inductive knowledge commonly believed to be supported by various forms of contextual interference. Top-performing students may be more likely to master classification without changes to the order of practice but still have room to improve in accurately carrying out a solution algorithm (e.g., procedural fluency). The academic performance of top students may therefore be more responsive to blocked practice if this form of practice is more efficient at promoting accuracy in carrying out an algorithm. This interpretation, while speculative, is consistent with existing meta-analysis results suggesting that the interleaving effect may be highly sensitive to the specific setting and learning material (Brunmair and Richter, 2019).

This study is the first evaluation of a whole-year interleaving program. Prior to our study, the duration of the longest interleaving mathematics program was four months (Rohrer et al., 2020b). The retention interval – the amount of time elapsed between the test and when the material was practiced – is also longer than previous studies. Although not observable directly, we calculate that the average item on the cumulative assessment was practiced approximately 141 days before the cumulative assessment. On a comparable measure, the longest interval studied previously was 92 days (Ziegler and Stern, 2014).[1] Our measure of long-term retention is a comprehensive assessment, including a mixture of material practiced nine months before the date of the assessment and some from weeks before the assessment.

Our study adds to an emerging literature that applies lessons from cognitive psychology to develop and evaluate scalable, low-cost educational interventions for LMICs (Dillon et al., 2017). Evidence of the interleaving impact on math skills comes from bespoke pilot studies designed with substantial researcher involvement and conducted in either the United States, Germany, or Switzerland (see Appendix Table A1). Despite the scientific importance of this literature, it cannot be assumed that increased interleaving in mathematics practice would increase learning in other settings given differences in populations, baseline learning levels, and potentially complementary inputs. Furthermore, impacts observed in pilots with substantial researcher involvement may not carry over to programs implemented at scale without researcher involvement by education practitioners (Al-Ubaydli et al., 2019; List, 2022). The present study confirms that a program designed and implemented by practitioners can reproduce the positive short-run impacts of pilot studies, but casts some doubt on whether these effects persist in the long run.

[1] Section 2.6 discusses details of how comparable retention intervals are calculated across studies.

While we find that short-term impacts may overstate longer-run gains, this does not mean that interleaving is not helpful. First, we cannot rule out moderate positive average impacts of interleaving on long-term retention. In many cases, manipulating the sequence of practice can be accomplished at a very low cost, so our results may still be consistent with the cost-effectiveness of some interleaving programs. Second, the substantial short-run impacts on test scores may be intrinsically important. In our study, the pace of instruction was held fixed for both blocked and interleaved conditions, but if interleaved practice supports faster learning, the pace of instruction could potentially be increased. Third, we find some suggestive evidence that interleaving impacts are more persistent at the bottom of the distribution suggesting that interleaving can guard against students falling behind. Fourth, the heterogeneous effects suggest that more effective practice might be developed by targeting based on students' skills.

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 summarizes our interpretation of these results and concludes.

2 Context

This section describes the context and details of both the blocked and interleaved practice programs. Section 2.1 describes the background and setting, including a discussion of the mathematics curriculum and in-class practice sessions. Sub-section 2.2 describes the how the interleaving program modified the practice sessions. Sub-section 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 problem set program took place in grade five math classrooms in 62 schools in Lagos state and Osun state. The schools in the study are all operated by Bridge Nigeria – a subsidiary of NewGlobe – which operates for-profit private schools located mostly 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 motivated by a concern within Bridge that some students were having difficulty classifying math problems (matching the appropriate solution concept to a given problem) on

cumulative assessments.[2] Bridge staff responsible for the central design of lesson materials consulted the cognitive science literature to identify ways to improve retention of mathematical knowledge, especially related to the ability to categorize mathematics problems and apply appropriate solution strategies.

Variation in exposure to interleaved practice was accomplished by modifying centrally standardized materials used to script classroom instruction. Bridge schools employ a model of structured pedagogy wherein teachers use digital guides that script nearly all classroom activities. The tablet-based lesson plan guides the teacher through the learning material step-by-step.

Each day students receive three periods of math instruction. Students are introduced to new material in two 45-minute periods. The two periods typically cover similar material, although there can be differences in the type of problem.[3]

The program modified practice embedded in classroom math instruction. During both 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"). All questions are written on the blackboard by the teacher based on instructions in the teacher guide. Sessions cover between six and fifteen problems focusing on a 'topic of the day.' Each topic is presented with an appropriate solution strategy. The next day, students move on to a new 'topic of the day' and complete practice problems related to that topic. There are two "my-turn/your turn" sessions in each math period, covering the same topic but different problems.

Outside of these mathematics courses, students have other opportunities for review. Weekly mathematics revision courses review material covered in the past week. Students are tested at mid- and endterm in mathematics, providing additional opportunities for retrieval practice (Karpicke and Roediger, 2008).

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.[4] The number and type of practice problems completed by the teacher during "my turn" were the same in both interleaved and blocked classrooms. The number of practice problems by students was also the same in blocked and interleaved classrooms, but the type differed. In comparison classrooms, "your turn" problems practiced by students were blocked so that each item used the same solution strategy aligned to the "my turn" segment. Interleaved "your turn" problem

[2]Students in Bridge schools are tested up to seven times per year in each subject: an initial diagnostic assessment followed by midterm and endterm assessments in each of the three terms in the academic year. Data from these tests are used centrally by Bridge to monitor the performance of lesson materials.

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

[4]In Term 1, only Mathematics 1 practice was blocked in both conditions. In Terms 2 and 3, classrooms assigned the interleaved condition received interleaved practice in both Mathematics daily sessions 1 and 2.

sets, on the other hand, contained both practice problems that were aligned with the topic of the day and practice problems that were not aligned. The aligned topics were drawn from the set of items practiced in the blocked practice sets. The non-aligned practice problems were drawn from the respective topics of the day covered in the ongoing academic term, and most material was drawn from the same unit (a set of interrelated topics that is typically covered in 3-5 consecutive lessons). There is variation across lessons in how the interleaved practice was formed. Most lessons included a mixture of approximately half aligned, and half non-aligned problems.[5] Figure 1 provides concrete examples of problems that students would receive in blocked and interleaved classrooms.

Schools receiving the interleaving program in grade 5 classrooms also received another novel program in grade 3. This program modified the level of difficulty of passage reading practice provided to students with initially more advanced reading skills. We think it is unlikely that the introduction of this program influenced the students in the interleaving study, although we cannot empirically test whether the grade 3 program influenced grade 5 student outcomes. Cross-grade-and-subject spillovers of minor pedagogical changes are rarely tested for by researchers.

2.3 Overall research agenda

The present evaluation is one of several such experimental evaluations of pedagogical programs done in collaboration with NewGlobe's Learning Innovation team. The Learning Innovation unit works to identify ways to improve learning in schools using NewGlobe's materials and test whether variations in materials are sufficiently effective to be implemented at a large scale. In addition, NewGlobe has worked 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 at the time of the program. One of these schools was excluded from randomization because it operated a slightly different model (known as "Bridge Plus"). The remaining 62 schools were randomly assigned to the interleaving condition or to the blocked problem sets condition. Each school contained one grade five classroom. We use the words school and classroom interchangeably to refer to the unit of randomization. In schools that received the interleaving program, Grade 3 classrooms also received a separate reading program that provided more difficult reading assignments to students with more advanced baseline literacy skills. While we are unable to rule out the possibility that the interleaving effect

[5]For additional details see Appendix A.

reflects the impact of this Grade 3 reading program, we think that such spillovers would be remarkable and without empirical precedent.[6]

Randomization was stratified using data on pre-assignment characteristics including lesson completion rates, student-teacher ratios, urban/rural location of the school, and a binary variable indicating whether the school had a grade five class in the previous year.[7][8] Figure 2 illustrates the construction of the final sample. Twelve strata were formed in total. The stratified randomization assigned 28 schools to the interleaved condition and 34 schools to the blocked condition. After randomization, it was discovered that two randomization strata contained only a single school. The randomization procedure deterministically assigned these strata to the comparison group, so these schools were dropped from the analysis.

We restrict the analysis sample both to ensure that estimated effects reflect causal effects of the program on students' learning and are not confounded by changes in the population of enrollees and to identify a sub-sample with high levels of follow-up to ensure that mitigate the risk of effects being compromised by selective attrition. The analysis includes those students who satisfy two criteria: 1) the student's unique identifier is listed in a data file for a mathematics test that NewGlobe administered immediately before the start of the program; 2) the data include pre-assignment information on the student's gender, age, and date of enrollment. In total, 525 students were excluded from the analysis based on these criteria. One school assigned to the interleaved group had no students meeting these criteria and was therefore dropped from the analysis. Note that these inclusion criteria do not appear to affect our main results.

2.5 Sample characteristics & baseline balance

We use student- and school-level data collected by Bridge.[9] Student-level data include gender, age, date of enrollment, and prior test scores. School-level data include a variable that classifies the location as either rural or urban (discussed below), a variable for the student-teacher ratio, and the number of years the school had been open.

The final study sample consists of 687 students (Table 1). The average age at baseline was 9.81 years and approximately half of the students are female. Students had been enrolled in a Bridge Nigeria school for 1.23

[6] A concern could be that introducing new versions of the lesson guides imposes an academy-level administrative cost which itself affects test scores. This study could underestimate the interleaving effect on test scores if the new program taxes the schools administrative capacity. However, we see this as a broader concern that affects a broad swath of experimental evaluations.

[7] Strata were formed by an indicator for whether a school had more than a 75 percent lesson completion rate (approximately the median), indicators for having 15 or fewer students per teacher, between 15 and 30 students per teacher, and greater than 30 students per teacher, and an indicator for whether the academy is classified as being in an urban location.

[8] The urban/rural data are Bridge's classifications used for setting cost-of-living salary adjustments.

[9] Readers may consider Bridge's incentives to accurately report data, given that they are a for-profit organization. The authors' view is that Bridge conducted this experiment with the intention of exploring whether a pedagogical intervention could improve educational outcomes of its students, and that the results, whether negative or positive, were not relevant to Bridge's reputation or profitability.

years on average when the interleaving program started.[10] The average student-teacher ratio (measured in the previous year) was 23. At baseline, the teacher guides logged the teachers completing scheduled lesson guides 77 percent of the time on average.

Most schools were in locations classified as rural areas (68 percent). Rural classifications are created by Bridge for administrative purposes. The classification is not based on objective criteria. Most schools are located in Lagos state, one of the most densely populated regions in sub-Saharan Africa. Rural locations in Lagos state may be more conventionally viewed as "peri-urban". Eight percent of schools are located in nearby, less urban, Osun state. All schools located in Osun are classified as rural.

Table 1 shows that both the student-level variables and the school-level variables are similar for the interleaved and blocked groups.

2.6 Test scores

NewGlobe developed termly posttests measuring short-term retention and cumulative assessment measuring long-term retention of material covered in the "my turn/your turn" sessions.[11] Each test was composed of thirty open-ended items that were graded 'correct' or 'incorrect' with no partial credit. Test items were sent to grade 5 mathematics teachers through the tablet. Teachers wrote the items on the board, and students recorded answers in their exercise books. Academy managers selected a teacher other than the grade 5 mathematics teacher to grade the tests. Students were given the correct answers after the test. For analysis raw scores are standardized to the comparison group mean and standard deviation. The timing of administration of the tests is shown in Figure 3.

The calculation of the retention interval in this study differs from other studies of the interleaving effect. The retention interval for an outcome is often calculated as the amount of time from the end of the interleaving intervention to the time of the test. In this study, tests were administered immediately after the end of a term of study, so that the retention intervals defined in this way would be only a few days. The present study covers material for an entire academic year, with some concepts having been practiced at the beginning of the program and others closer to the test. We propose an alternative measure of the retention interval which can be easily calculated for prior studies that captures the fact that the amount of time elapsed since the practice event for different concepts may vary within a test. Specifically, we calculate the retention interval as the average duration (in days) from the beginning and the end of a practice period (either an academic term in the case of the posttests or the academic year for the cumulative assessment) and the date of the

[10] The data indicate the enrollment date in Bridge Nigeria schools. If a student changed schools, this date would be the earliest date they enrolled in any Bridge Nigeria school.

[11] These tests were administered in addition to Bridge Nigeria's regular mid- and end-term summative assessments.

test.[12]

Posttests (short-term) Posttests measure short-term retention of the specific topics covered in the problem sets for each term. Most subjects were included in at least two terms, and some, such as fractions and arithmetic, were covered in all three terms (Table 2). However, the specific problems and their corresponding solution concepts were largely distinct across terms. For example, an arithmetic question on the Term 1 posttest involves addition and subtraction of four-digit numbers, multiplication of three-digit numbers, and division of a three-digit number by a single-digit number. In Term 2, arithmetic topics include subtraction of numbers with one decimal place, story problems requiring multiplication of velocity and time, and division of numbers with decimal places.

We use the phrase "short-term retention" to distinguish these outcomes from the cumulative assessment. The average retention interval of the posttests is 36 days. The retention interval of the individual posttests varies due to differences in the length of academic terms. The retention interval of the first posttest is 45.5 days, the second is 38.5, and the third is 23.[13]

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

Pretests mirrored the posttests, capturing knowledge of material before it was covered in instruction and problem sets. Pretests and posttests administered in the same academic term covered the exact same topics but with minor alterations to each item. For example, a pretest item was: 'Solve 9857 - 1903 - 708', and the accompanying posttest item was 'Solve 9056 - 2071 - 609'. The order of topics covered differed for pretests and posttests.

Cumulative test score (long-term) The cumulative assessment was administered at the end of the third academic term, just after completing the interleaving program. The cumulative test covered the content of the lessons over the full academic year with one-third of the questions covering material from each of the

[12]Specifically, if students practiced a set of concepts $\mathcal{X}_j$ across a time interval from t_{0j} to t_{1j}, then the average retention interval for a test y_j administered at time t_{3j} is calculated as $\frac{\Delta_{1j}+\Delta_{2j}}{2}$, where $\Delta_{zj} = t_{3j} - t_{zj}$.

[13]We note that the retention interval of the posttests is longer than what is conventionally referred to as "short-term" in the interleaving literature, which often means impacts in the same day (e.g., Patel et al., 2016), or week (e.g., Taylor, 2008; Taylor and Rohrer, 2010). Relative to these studies, our short-term results may be viewed as reflecting longer-term retention. However, we note that the retention interval of these assessments is similar to other large field experiments of the interleaving effect and prior studies examining the persistence of effects over intervals as short as the posttests find, if anything, that very short-run impacts tend to be smaller than longer-run impacts (Brunmair and Richter, 2019).

[14]Collapsing multiple outcomes into one in this way has the advantage of conciseness, higher follow-up rates (because analysis is conducted on the set of students with at least one posttest result), and 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. Appendix Figure A1 show that the average impacts are similar across the three posttests.

three terms. The cumulative assessment was deliberately designed to include the same problem types covered in the previous three termly posttests described above. Thirty percent of the items from the cumulative test are analogous to items found in the Term 1 posttest, 37 percent are found in the Term 2 posttest, and 33 percent are found in the Term 1 test (Table 2). The retention interval for the cumulative assessment, calculated using the method described above, is 141 days.

Approximately 78% of students in comparison schools have cumulative test data. 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 among students in the interleaved and blocked conditions may raise concerns that the results are compromised by selective attrition. While we cannot rule out this possibility, empirical evidence weighs against this hypothesis. Because the effect of interleaved practice on individual posttests is similar (Appendix Figure A1), and most students have at least one test over the course of the program, we do not think that the estimated effects are compromised by attrition. Furthermore, (non-) attrited students do not appear to be higher- or lower-performing at baseline (Appendix Table A3). We find that lower follow-up rates among students in interleaved classrooms are partly explained by student absence from class on the day of the test (Appendix Table A4). Attrition cannot be explained by teachers not administering the test or that students are missing from the test file. One interpretation of the impact on attrition on individual tests could be that students find interleaved practice more challenging (Taylor, 2008; Ziegler and Stern, 2014, e.g.), become discouraged and attend school less frequently.[15]

[15] Pupils were not informed in advance about the tests, and so it is unlikely that they were avoiding the assessment itself.

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. To do so, we estimate the following linear model of test scores

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

where $Z_j \in 0, 1$ indicates 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] Estimating Equation 1 yields $\hat{\beta}_s$. 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 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 mean zero in the blocked practice group. We use the mean of the three midterm and endterm math scores prior to the start of the program. 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.

[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 (Abdulkadiroğlu et al., 2017; Angrist et al., 2022; Borusyak and Hull, 2020) and conserves the sample size for estimation. Results are broadly similar when controlling for strata fixed effects.

[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 A2). 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).

Second, we use quantile regression to examine distributional impacts further. Estimation of Equation 2 can understate the degree to which the interleaving effect varies with baseline math skills if the scores contain measurement error.[18] Quantile regression gives the interleaving effect $\hat{\beta}_s^\tau at$ a chosen percentile τ of the test score distribution. Under an assumption that the relative ranks of children are preserved regardless of the form of practice, quantile regression estimates can be interpreted as giving the effect *on* students at different initial rankings within the classroom. In our case, we take the similarity in the results from estimating the interaction and the quantile regression as evidence that rank-preservation holds approximately.

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.

4 Results

This section shows that interleaved practice in mathematics class yields only small positive and not statistically significant increases in the longer-term cumulative assessment test scores compared to blocked practice. Interleaving yields larger gains on shorter-term test scores and the effect is largest for students at the bottom of the test score distribution.

Main results The short-term effect of the interleaved practice on the posttest measuring end-of-term retention is 0.29 standard deviations (Table 3, Panel B, Column 1). This effect would be equivalent to moving a student from the median of the comparison group distribution of test scores to the 61^{st} percentile. This effect size is very close to that obtained from a meta-analysis of evaluations comparing the short-term effects of interleaving to blocked practice in mathematics education (0.34 standard deviations) (Brunmair and Richter, 2019). We find that the impacts on individual posttests are very similar across the three posttests (Appendix Figure A1). It is perhaps noteworthy that the largest impact comes from the third posttest which also happens to be the test with the shortest average retention interval (23 days compared to 45.5 and 38.5 for the first and second posttests respectively). However, the difference between the interleaving effect on the third posttest and the other posttests is not statistically significant.

[18] Quantile regression estimates come from the same linear model as Equation 1. Quantile regression estimates come from minimizing the loss function as expressed below:

$$(\alpha_s^\tau \ \beta_s^\tau \ \gamma_s^\tau)' = \arg\min_{(\alpha \ \beta \ \gamma)' \in \mathcal{R}^3} \sum_{i=1}^{n} \left(\rho_\tau \left(Y_i js - \alpha_s - \beta_s Z_j - \gamma_s p_j \theta\right)\right), \tag{3}$$

where $\rho_\tau(x) = x(\tau - \mathbb{1}\{x < 0\})$.

The effect of interleaved practice on the cumulative test score is 0.04 standard deviations (Table 3, Panel B, Column 2). The null hypothesis that the impact of interleaving on the cumulative assessment is zero cannot be rejected. Given the standard error, we are unable to rule out moderate positive or negative impacts of interleaving on the cumulative assessment. The 95 percent confidence interval includes effects from -0.27 to 0.35 standard deviations.

The longer-term effects are smaller than the short-term effects. A test of the hypothesis that the effects on the cumulative assessment and posttests are equal can be rejected at the 5 percent level (Table 3, Panel B). Given the large positive impacts on the Term 3 posttest that was administered only days before the cumulative assessment, and the fact that one-third of the items on the cumulative assessment were from the Term 3 posttest, which was administered only days before the cumulative assessment, these results suggest that the impacts on retention of material from the first two terms may be even lower. If we assume that (a) the overlapping material on the cumulative and Term 3 assessments were a representative subsample of the Term 3 posttest items, and (b) that the impact on the material in the Term 3 posttest persisted to the date of the cumulative assessment, then this implies that the point estimate of the implied effect on Term 1 and 2 material would be -0.1 standard deviations (CI: -.47 - .27).[19]

Distributional effects Interleaving appears to have larger, more persistent, and more robust impacts on lower-performing students than on higher-performing students. [20] 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.

Estimates of the interaction between interleaved practice and baseline test scores reveal negative point estimates across both tests (Table 3, Panel C), although only in the case of the short-term posttest is this interaction statistically significant. For students at the mean of the test score distribution, the interleaving effect on the posttest is estimated to be 0.35 standard deviations (Column 1). For a student with a one standard deviation greater baseline test score, the effect is 0.19 standard deviations lower, or 0.16 standard deviations. Alternatively, for a student who is one standard deviation below the mean, the interleaving effect is estimated to be 0.54 standard deviations. For students at the mean of the test score distribution, 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 perhaps noteworthy

[19] This estimate is obtained by noting that the effect on the cumulative assessment $\beta_{\text{cumulative}} = 0.37\beta_{\text{term 1}} + 0.30\beta_{\text{term 2}} + 0.33\beta_{\text{term 3}}$ given the composition of the cumulative assessment shown in Table 2 and the assumptions described above.

[20] We also test for heterogeneous impacts of interleaved practice by gender (Table A6). Average treatment effects on the posttests are larger for male pupils, but we find no statistically significant differences in treatment effects for male and female pupils.

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 percent are respectively 0.49, 0.44, 0.44, and 0.38 standard deviations and statistically significant.[21] The estimated effect on the 20th percentile is 0.49 standard deviations and highly statistically significant (Figure 5) on the long-term test.[22]. Notably, the estimated effect at the 90th percentile is -0.33 standard deviations and we are able to reject the null hypothesis at the five percent level, suggesting that interleaving may be harmful for top performers.

Adjusting the cumulative assessment to account for the material from Term 3 suggests that the impact on long-term retention may have been even smaller. Figure 5 shows two approaches to partial out performance on the Term 3 posttest from the cumulative assessment. The bottom left panel reports 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 including the Term 3 posttest score as a control. 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.

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

5 Discussion and conclusion

This study evaluates a nine-month – full academic year – program that increased the amount of interleaved practice used in grade 5 mathematics classwork in Nigeria. Students in the interleaved program would be exposed daily to practice problems related to the material covered in that term so far, with about half related to the topic of the day and the other half related to the previous topics of the day from that same term. Students in blocked comparison classes would practice the same number of problems, but the material was always related to the topic of the day. Our analysis finds that interleaving had statistically significant, and positive impacts on short-term retention tests covering material practiced for each of the prior three respective school terms. These impacts are similar to those observed in the prior literature from high-income countries. However, on an end-of-year cumulative test covering materials from across the three

[21] For a visual comparison of the test score distributions see Figure 4.
[22] Appendix Table A7 reports results in table form.

terms, students in the interleaved program had scores that were statistically indistinguishable from those comparison students who had been following blocked practices. The absence of an aggregate impact may reflect heterogeneous distributional impacts of interleaving: while the bottom of the long-term test score distribution appears to have improved from interleaved practice, these gains are offset by negative impacts at the top.

The pattern of effects observed can be explained by recognizing that math learning combines many inter-related skills (e.g. Silver, 1986; Rittle-Johnson et al., 2001). While interleaving may support learning about categories (e.g., matching problems to solutions) it may be a less efficient way to develop other mathematical skills, such as the ability to carry out the solution without making errors (i.e., "procedural knowledge").[23] Low-achieving students may struggle with the first step of linking problems with solution strategies, and therefore benefit more from interleaved practice. A higher-achieving student might be better able to master the classification step over time and therefore would benefit more from practice that helps them develop procedural fluency.

There are several reasons why higher-achieving students might need less support learning to classify problems. First, top math students in the present study tend to perform better even before math practice, as measured by pretests. This suggests they may have higher levels of conceptual knowledge which could help them acquire foundational classification skills more rapidly.[24] Second, there may be other opportunities for effortful retrieval that may benefit students with initially stronger math skills. Regular testing and review modules give students distributed practice at longer retrieval intervals. Lower-achieving students may benefit less from these retrieval opportunities if they have less ability to remember material from one representation to the next. Third, behavioral responses to the program may also explain differences in learning needs (Todd and Wolpin, 2003). Top-performing students may be more likely to use self-study to address gaps in their ability to recognize different types of problems.

Heterogeneous impacts of the form of practice have important policy implications. First, the standard practice of blocking practice early on in instruction may be less suitable for struggling students. The prevalence of blocked practice (e.g. Rohrer et al., 2020a) may reflect historical biases toward supporting the development of the most advanced students (Kremer, 2003; Glewwe et al., 2009). Increasing interleaving may be especially important in low- and middle-income countries where low learning levels are common (Patel and Sandefur, 2020; Angrist et al., 2020). Second, one-size-fits-all practice sequences may be sub-optimal in some cases. Several studies have demonstrated that adapting the level of difficulty of instruction to a

[23] Interleaving is believed to support inductive learning where students must learn to discriminate between two categories, especially where the differences between categories are subtle (Brunmair and Richter, 2019), but evidence is mixed for other learning tasks (e.g. Noh et al., 2016).

[24] Pretests are sometimes used to measure conceptual knowledge (e.g. Rittle-Johnson et al., 2001).

student's level can produce large learning gains (Banerjee et al., 2007; Duflo et al., 2011; Banerjee et al., 2017; Muralidharan et al., 2019). The results of the present study suggest that, in addition to the level of instruction, differentiating the form of practice depending on students needs may also produce larger learning gains on average. Future work is needed to assess whether differentiation along these lines can produce benefits.

Large short-term impacts on lower achieving students may themselves be policy-relevant. First, students who struggle with math early may be more likely to develop math anxiety (Ramirez et al., 2013) or a fixed mindset (Blackwell et al., 2007), which could hold back effort in math later in life. Second, larger short-term impacts on retention of students more likely to struggle with newly taught math skills may allow for adjustments to the pace of instruction. In the present study, the amount of material was held fixed, but in settings where the pace of instruction is slowed to accommodate lower-achieving students, interleaved practice allows for a faster pace including advanced material that may benefit higher-achieving students (Cohodes, 2020).

Broadly, our results suggest that using cognitive science to develop effective policies may require consideration of counterfactual opportunities to encode long-term retention of mathematics material. Scientific studies of memory typically control counterfactual learning conditions to establish clear causal connections between variables of interest. However, students learn in complex environments with multiple opportunities to rehearse their skills. For example, the three posttests used to measure short-term retention themselves provided an opportunity for retrieval practice (Karpicke and Roediger, 2008) that was itself interleaved, potentially diminishing the contrast between the two conditions. Furthermore, students are not passive actors in their learning: the net effect of any educational policy reflects the direct effect of the policy as well as the impacts on the behavior of students, their parents, and their teachers (Todd and Wolpin, 2003).

We note some alternative explanations for the fadeout of the effect and why we feel that these are less compelling, but which would have different policy implications.

Broadly, there are two possible reasons why the interleaving effect would fade out: forgetting in interleaved classrooms and catch-up in blocked classrooms. While we cannot rule out forgetting, there are several reasons to suspect its role was small. The short-term assessments reflect skills learned over a relatively long period of time (36 days on average) and indicate long-term memory consolidation as opposed to short-term test cramming. The broad similarity of math topics in this setting and frequent testing indicate high levels of repetition and retrieval practice, so it's unclear why newly learned skills, consolidated over a month, should then be rapidly forgotten. Also, students in both the interleaved and blocked conditions answered more items correctly on the cumulative assessment than either the Term 2 or Term 3 assessments (see Appendix Figure A3).

Another unlikely explanation is that the absence of an effect on 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 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). Also, this interpretation does not easily explain the negative impacts at the top of the distribution on the cumulative assessment.

A final interpretation for which we find little evidence is that the cumulative assessment is 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. However, as we note in our discussion of attrition, we see no evidence of selection on observable characteristics conditional on follow-up on the cumulative assessment, as shown in Appendix Table A3. 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 (Appendix Figure A1). Finally, Appendix Table A3 reports covariate balance between the interleaved and blocked condition conditional on follow-up on all of the individual posttests. On observable characteristics, we see no evidence of differences in the composition of the sample conditional on follow-up for any single test.

References

Abdulkadiroğlu, A., J. D. Angrist, Y. Narita, and P. A. Pathak (2017): "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 85(5), 1373–1432.

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., G. Gray-Lobe, C. M. Idoux, and P. A. Pathak (2022): "Still Worth the Trip? School Busing Effects in Boston and New York," Working Paper 30308, National Bureau of Economic Research.

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

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.

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.

Borusyak, K. and P. Hull (2020): "Non-Random Exposure to Exogenous Shocks: Theory and Applications," Working Paper 27845, National Bureau of Economic Research.

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.

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.

Glewwe, P., M. Kremer, and S. Moulin (2009): "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, 1, 112–35.

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.

Kremer, M. (2003): "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons," *The American Economic Review*, 93, 102–106.

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

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.

Noh, S., V. Yan, R. Bjork, and W. Maddox (2016): "Optimal sequencing during category learning: Testing a dual-learning systems perspective," *Cognition*, 155, 23–29.

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, D. and J. Sandefur (2020): "A Rosetta Stone for Human Capital," *CGD Working Paper*.

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

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.

Rittle-Johnson, B., R. Siegler, and M. Alibali (2001): "Developing conceptual understanding and procedural skill in mathematics: An iterative process." *Journal of Educational Psychology*, 93, 346–362.

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

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

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

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

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

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.

Silver, E. (1986): "Using conceptual and procedural knowledge: A focus on relationships," *En J. Hiebert (Ed.), Conceptual and procedural knowledege: The case of mathematics*.

Taylor, K. and D. Rohrer (2010): "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.

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.

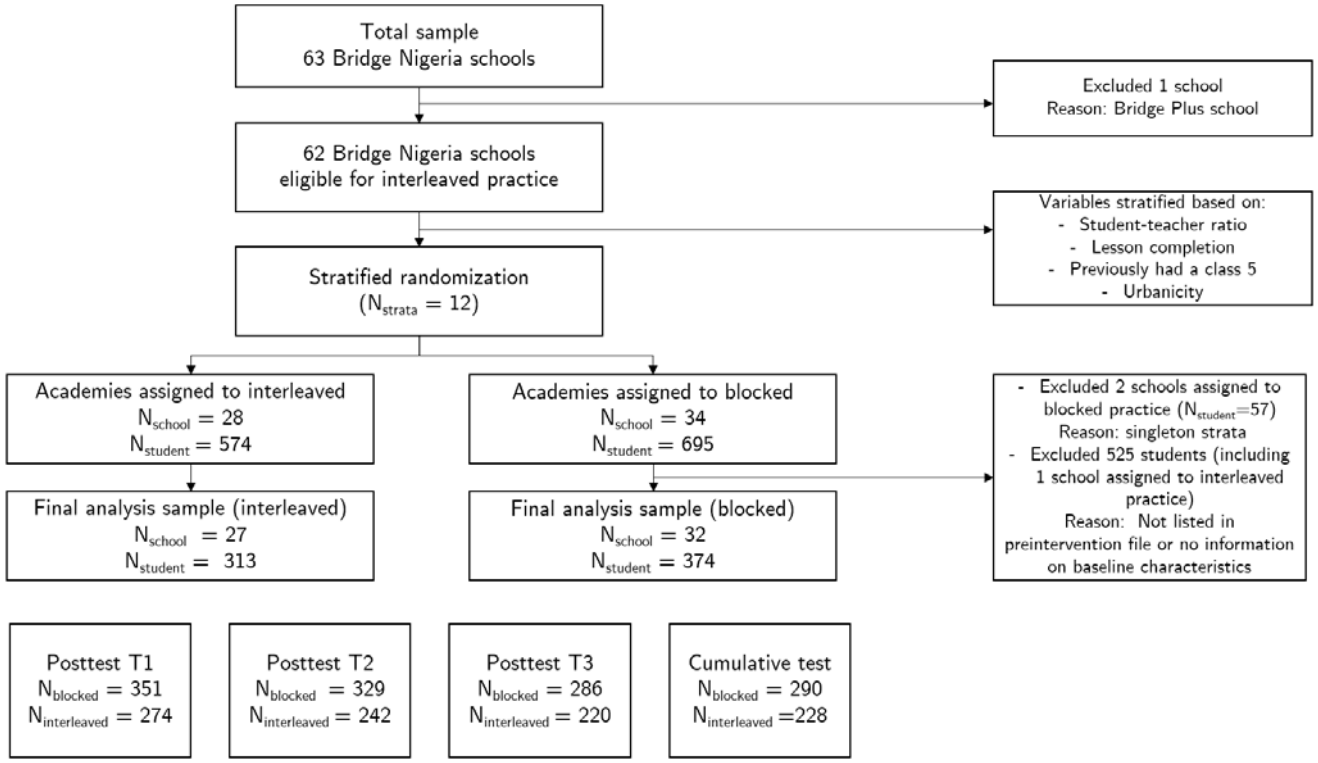
Figures & Tables

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?

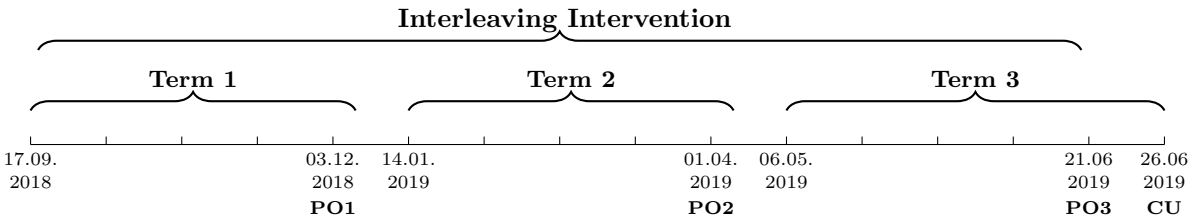
Notes: Examples of blocked practice vs. interleaved practice (independent study). All items in the blocked practice set cover the topic of the day. Items in the interleaved practice set cover the topic of the day (problem 1, 3, 5, 8), topics from previous lessons (problem 2, 4, 6) and an integrated problem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

Figure 2: Consort diagram



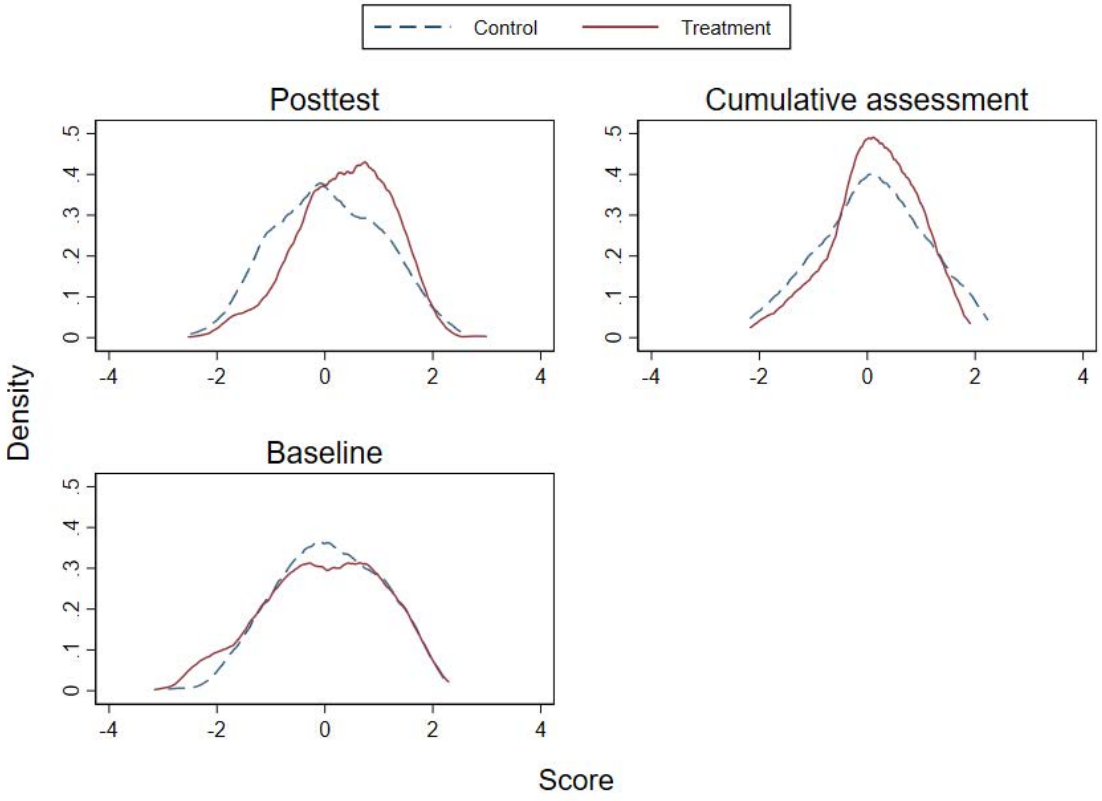
Notes: Figure shows the process of arriving at the final sample of schools and students for analysis. The final row shows the number of students in blocked and interleaved schools that have a test for each of the tests.

Figure 3: Timeline of events



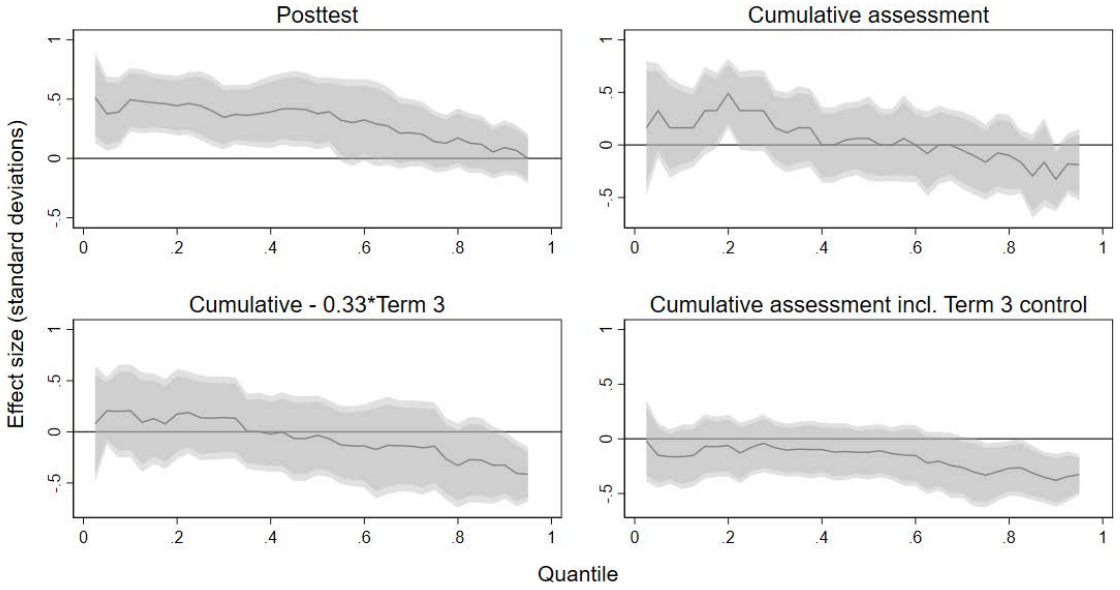
Notes: Figure shows the project timeline. PO=Posttest and CU=Cumulative assessment.

Figure 4: Test score distributions



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

Figure 5: Effects across quantiles



Notes: Figure shows the effects of interleaving at percentiles of the distribution of test scores. The line represents the point estimate at each ventile. The darker area represents the 10% confidence interval and the lighter area represents the 5% confidence interval.

Table 1: Sample characteristics at baseline

	Mean (1)	Blocked Mean (2)	Interleaved Mean (3)	P-value (4)	Number of students (5)	Number of schools (6)
Panel A: Student-level characteristics						
Age	9.81	9.79	9.83	0.65	687	59
Years enrolled	1.23	1.20	1.28	0.60	687	59
Female	0.51	0.48	0.54	0.13	687	59
Baseline score	0.02	0.07	-0.05	0.41	476	43
Panel B: School-level characteristics						
Years the school is open	1.53	1.46	1.60	0.43	687	59
Student-teacher ratio	23.27	23.00	23.59	0.93	687	59
Average percentage of lessons completed	0.77	0.77	0.77	0.69	687	59
Rural (v.s. Urban)	0.68	0.69	0.67	0.87	687	59
Osun state	0.08	0.09	0.07	0.79	687	59

Notes: P-values reported are from regressing the baseline characteristics on treatment (t-statistic), controlling linearly for the probability of the schools' assignment to the interleaving condition. For the individual-level variables the standard errors are clustered by school. Student-teacher ratio and average percentage of lessons completed are based on values from previous years. The Female-, Rural-, and Osun row report p-values from Z-tests. F(academy-level variables) = 0.34, F(individual level variables) = 0.44. ***, **, 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

	Posttest (1)	Cumulative (2)
Panel A: The effect of interleaved practice on follow-up		
Interleaved practice	0.01	-0.04
	(0.01)	(0.04)
Blocked practice mean	0.97	0.78
Number of students	687	687
Number of schools	59	59
Panel B: The effect of interleaved practice on test scores		
Interleaved practice	0.29**	0.04
	(0.12)	(0.16)
P-value that effect is equal to effect on cumulative test	0.02	
Number of students	665	518
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 x Baseline testscore	-0.19**	-0.05
	(0.09)	(0.16)
Baseline testscore	0.70***	0.55***
	(0.08)	(0.12)
Number of students	457	362
Number of schools	43	43

Notes: Baseline test score, in Panel C, is the average of 6 standardized baseline math scores (3 midterms and 3 endterms). We control linearly for the probability of the schools' assignment to the interleaving condition in all specifications. Standard errors are clustered at the school 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 units 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 $L(n-z)$

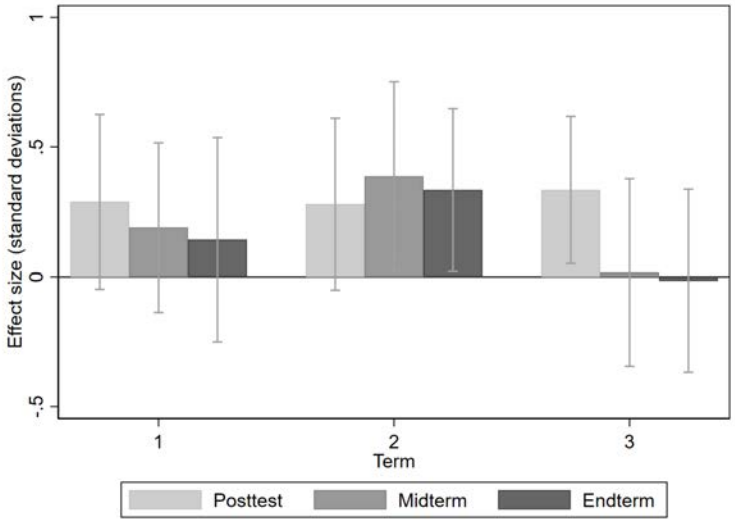
The penultimate question $L(n-z)$ was selected to introduce the final, integrated problem. It always was placed at the end of the sequence of practice.

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.

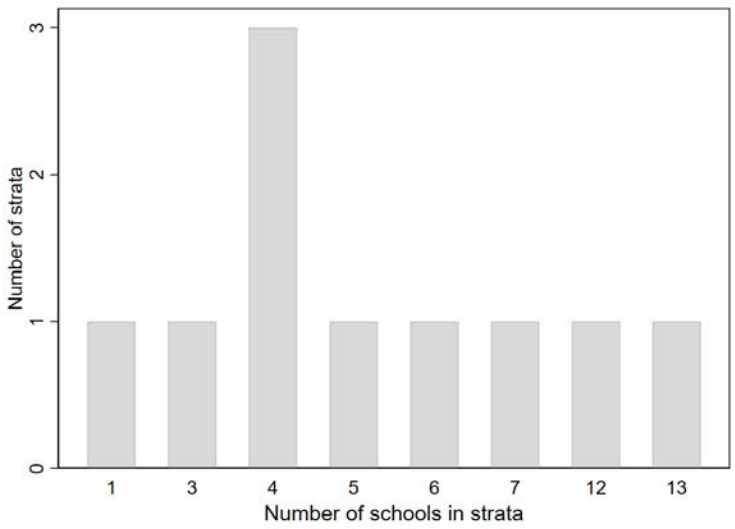
Appendix B Appendix Figures and Tables

Figure A1: The effects of interleaved practice on math test scores by term



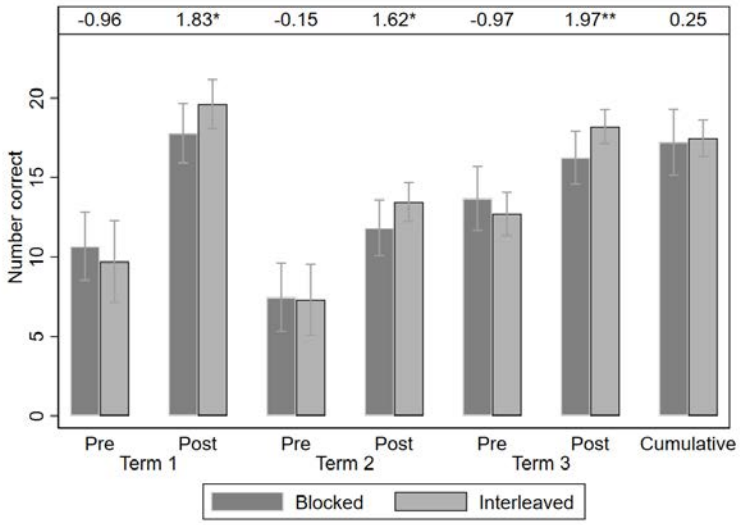
Notes: This figure shows the effect of interleaving compared to blocked problem sets on test scores in each academic term.

Figure A2: 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 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%.

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 (2010)	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
Bridge Nigeria (this study)	687	59	3	36	0.29	Nigeria
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 study (this study)	687	59	9	141	0.04	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 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

	Posttest follow-up (1)	Posttest test scores (2)
Interleaved practice	-0.05*	0.30**
	(0.02)	(0.12)
Blocked practice mean	0.86	0.07
Number of students	2,061	665
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. In both specifications, we control linearly for the probability of the schools' assignment to the interleaving condition interacted with a 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

	Posttest sample (1)	Cumulative test sample (2)	Posttest t1 sample (3)	Posttest t2 sample (4)	Posttest t3 sample (5)
Age	0.05	0.05	0.05	0.04	0.01
	(0.09)	(0.10)	(0.09)	(0.09)	(0.09)
Female	0.06	0.06	0.08*	0.08*	0.08
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
Years enrolled	0.07	0.05	0.09	0.04	0.05
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
Baseline test score	-0.14	-0.08	-0.16	-0.16	-0.10
	(0.17)	(0.17)	(0.17)	(0.18)	(0.17)
Number of students	665	518	646	582	506
Number of schools	58	58	58	57	58
Number of students with baseline test score	457	362	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 the cumulative test (Column 1) or on a posttest (Column 2-5). 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. In all specifications we control linearly for the probability of the schools' assignment to the interleaving condition. Standard errors are reported in parentheses and are clustered at the school level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A4: The effect of interleaved practice on additional measures of follow-up rates

	Posttest term 1 (1)	Posttest term 2 (2)	Posttest term 3 (3)	Cumulative test (4)
Panel A: The effect of interleaved practice on (pupil) being marked absent				
Interleaved practice	0.00	0.02	0.06	0.04
	(0.02)	(0.03)	(0.05)	(0.04)
Blocked practice mean	0.06	0.12	0.24	0.22
Number of students	683	686	686	687
Number of schools	59	59	59	59
Panel B: The effect of interleaved practice on assessment not administered				
Interleaved practice	(-)	0.09	(-)	(-)
		(0.06)		
Blocked practice mean		0.00		
Number of students	683	686	686	687
Number of schools	59	59	59	59
Panel C: The effect of interleaved practice on pupil's absence from test file				
Interleaved practice	0.007	-0.003	0.003	(-)
	(0.006)	(0.003)	(0.003)	
Blocked practice mean	0.003	0.003	0.000	
Number of students	687	687	687	687
Number of schools	59	59	59	59

Notes: Standard errors are clustered at the strata 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. We control linearly for the probability of the schools' assignment to the interleaving condition in all specifications.

Table A5: The effect of interleaved practice on math test scores, using the Wild Cluster Bootstrap-t procedure

	Posttest (1)	Cumulative (2)
Interleaved practice	0.29***	0.04
Wild bootstrap *pvalue*	0.04	0.87
Wild bootstrap CI	[0.015, 0.637]	[-0.421, 0.458]
Number of students	665	518
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	0.40**	0.17
	(0.16)	(0.20)
Interleaved practice x Female	-0.20	-0.25
	(0.14)	(0.20)
Female	0.03	0.11
	(0.11)	(0.16)
Interleaving effect for female students	0.20*	-0.08
	(0.12)	(0.17)
Number of students	665	518
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. 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

	Posttest (1)	Cumulative (2)
10th percentile	0.49***	0.16
	(0.14)	(0.21)
20th percentile	0.44***	0.49***
	(0.13)	(0.17)
25th percentile	0.44***	0.33*
	(0.15)	(0.20)
50th percentile	0.38**	-0.00
	(0.15)	(0.19)
75th percentile	0.14	-0.16
	(0.13)	(0.18)
80th percentile	0.17	-0.10
	(0.13)	(0.20)
90th percentile	0.09	-0.33**
	(0.12)	(0.15)
Number of students	665	518
Number of schools	58	58

Notes: Each estimated effect is from a separate quantile regression. In all specifications, we control linearly for the probability of the schools' assignment to the interleaving condition. Standard errors are clustered at the strata 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.