Details, Delegation, and Control*
The Impact of Streamlining Instructions to Teachers on Student Test Scores

Guthrie Gray-Lobe† Michael Kremer‡ Joost de Laat§ Wendy Wong¶

PLEASE DO NOT CIRCULATE
February 6, 2025

Advances in information technology have expanded managerial span of control in part by allowing for easier dissemination of detailed instructions to front-line workers in service sectors (Hsieh and Rossi-Hansberg, 2023). Gray-Lobe et al. (2022) estimate large learning gains from applying this model to education in an LMIC. This paper analyzes two experiments by a Kenya-based private school provider, showing that simplifying lesson plans to allow more teacher discretion improved student test scores by 0.17 standard deviations, with the largest gains in larger classrooms. Less detailed guides reduced score dispersion within classrooms. We find evidence of positive impacts on teacher attendance, suggestive of higher levels of job satisfaction. We examine impacts on the distribution of classroom effects within a simple model of standardization wherein explicitly codifying how a task should be performed places a floor on a task's performance at the expense of placing a ceiling. Consistent with this, the second experiment showed more dispersed classroom effects with more high-performing classrooms. However, the first experiment found more compressed effects with large gains for the bottom of the distribution. These results suggest that detailed instructions may impose cognitive load, particularly in lower-performing classrooms.

*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 New-Globe Education. We thank Shannon May, Sean Geraghty, Tim Sullivan, and the participants of the 2024 Midwest International Economic Development Conference for valuable feedback.

†University of Chicago, email: graylobe@uchicago.edu
‡University of Chicago, email: kremer.m@gmail.com
§Utrecht University, email: j.j.delaat@uu.nl
¶University of Chicago, email: wendywong@uchicago.edu

1 Introduction

A large literature has shown how education can be improved by providing teachers with instructions through teacher guides designed by centralized lesson planners (Piper et al., 2015, 2018; Romero et al., 2019; Fazzio et al., 2021; Gray-Lobe et al., 2022). Developing an effective lesson plan is a complex task, requiring knowledge of mnemonic strategies and other instructional best practices. Teachers may lack the time and knowledge to design a good lesson, especially if knowledge of subject matter and effective pedagogy is low (Bold et al., 2017). Structured pedagogy divides the task of teaching into lesson plan development and lesson plan delivery sub-tasks, separating the execution of these sub-tasks across workers according to their comparative advantages. This division of labor has been found to improve student test scores in both rich (Stockard et al., 2018) and poor countries (Akyeampong et al., 2023).

In designing teacher guides, a manager must consider the amount of detail to give. Some tasks may be better performed by a teacher, even if there are others for which codified instructions would enhance quality. In determining the optimal level of detail, managers must consider teachers' skills and knowledge, including local information that might not be available to the designer (Holmstrom, 1980), such as student needs. Too much detail can have adverse effects. Teachers may be better positioned to do tasks like pacing instruction and choosing their phrasings as local information emerges, even if they have a disadvantage in terms of other pedagogical tasks like sequencing material and selecting example problems.

As the use of structured pedagogy has expanded, so have concerns that excessive scripting may harm learning if teachers cannot adapt lessons to local classroom needs (Dresser, 2012). Surveying 16 studies from 13 countries, Piper et al. (2018) shows that less detailed teacher guides are associated with larger impacts on student learning. This study complements these findings with causal RCT evidence measuring the impacts of edits to teacher guides intended to reduce the level of detail.

The study took place in Kenyan private schools operated by a firm that uses centrally-designed teacher guides to script nearly all class instruction. Each day, guides are disseminated to thousands of teachers using a tablet. Teachers are specialized in delivering classroom instruction following the steps provided in the teacher guides.

Two separate experiments compared the usual status quo guides to those edited down to be

less detailed. The resulting teacher guides omitted many details that explicitly stated how a topic should be explained. Schools were randomly assigned to receive either the status quo teacher guide or the alternative, less-detailed version. The overall length of each class and the curriculum material covered were the same for both groups. Less-detailed guides increased test scores by 0.17 standard deviations by the end of the trials.

Class size mediates the effects in both experiments. The average test score impact is largest in those classrooms with the largest class sizes. Within classroom residual test scores are also less dispersed in classrooms with above-median size.

A simple model of task standardization predicts that instituting codified instructions for how to complete a task should reduce the dispersion in that task. Importantly, explicit instructions regarding how a task should be performed will, if adhered to, put a floor on the quality of that task. Removing instructions should lead to an increase in the dispersion in class effects (or value-added) and the collapse of that floor.

While we find some evidence for this simple model in the second experiment, where less-detailed guides are associated with an increase in the number of high-performing classrooms, less-detailed guides are associated with fewer very low-performing classrooms in both experiments.

These results suggest that more-detailed guides can waste other scarce resources in classrooms, such as teacher cognitive bandwidth or time. Reading more detailed instructions requires mental effort and the use of working memory and may also slow the teachers progress. Literature in psychology has demonstrated how cognitive strain can impede communication skills including both the abilities to give instructions (Kurinec et al., 2019; Roßnagel, 2000, 2004; Kurinec et al., 2019), follow instructions (Yang et al., 2015), and receive feedback (Roßnagel, 2004), three skills that are essential to effective teaching. The reduction in the fraction of classrooms that severely underperform from *reducing* the amount of detail in instructions is consistent with a model in which low-performing teachers are especially sensitive to this cognitive strain effect. Low-performing teachers may be those who enter class less prepared or those who are more reliant on the guide due to gaps in prior knowledge. Cognitive load effects may also explain the robust interaction with class size. In other settings, larger class sizes are often associated with more disruptive classrooms and behavioral problems to manage (Lazear, 1999), less personal connections to students (Finn et al., 2003), and higher levels of reported teacher stress (French, 1993; Huang et al., 2022). Teachers in

large classrooms may therefore have less bandwidth to communicate effectively.

Another interpretation is that less-detailed guides may save teachers time to spend on individual feedback or uncodified classroom management tasks. If working through more detailed instructions consumes more time, teachers have more time to provide individual feedback. If there are diminishing returns to individual feedback, this benefit would tend to be greatest in classrooms with more students.

Providing the appropriate amount of detail in instructions may also improve worker satisfaction. Teachers who received less-detailed guides took fewer days off in the second experiment. Estimates in the second experiment are smaller and not significantly different from zero. While we can only speculate as to the mechanism driving the effect in the second experiment, we consider three possible (non-exclusive) interpretations. First, teachers may enjoy greater discretion in choosing how they explain topics, even when the scope for choice is limited. Second, they may have felt that the guides were easier or more satisfying to use. At least in the second experiment, teachers were slightly more likely to complete the lessons. The latter interpretation suggests that the results may be informative about the role of information overload and "technostress" (stress experienced by end-users of information technology) on worker strain and burnout (Ragu-Nathan et al., 2008; Arnold et al., 2023). Third, teachers may have derived greater satisfaction from more rapid progress of students in those classrooms with less-detailed guides.

These studies were not designed to evaluate the overall effectiveness of teacher guides themselves, as they did not create variation in the amount of content for which instruction was scripted.[1] A large body of work shows that teacher guides can be an effective intervention in schools (Piper et al., 2018; Romero et al., 2019; Fazzio et al., 2021). The status quo materials used in comparison classrooms were designed by experts with training from leading graduate schools of education. Gray-Lobe et al. (2022) provide experimental evidence that the particular model employed in this study raises timely grade progression by 20 percentage points and an index of math and language scores of primary school-aged children by an amount equivalent to 0.89 equivalent years of business-as-usual school (EYS) compared to other local schools. The results of this study illustrate how already effective instructions to teachers can be improved marginally through editing.

[1] In both experiments both versions of the guides covered the same set of topics and practice problems. The second experiment arguably allowed for marginally more teacher autonomy by reducing the degree to which instruction for each topic was scripted.

Our findings are consistent with the observation in Piper et al. (2018) that the impacts of structured pedagogy programs promoting foundational skills are diminished when guides are more detailed. The correlation in Piper et al. (2018) could reflect unobserved characteristics of the interventions themselves if, for example, longer guides are used in settings where gaps in teacher knowledge are greater. Our study offers the first experimental estimates of the impact of reducing the level of detail on the effectiveness of teacher guides. [2]

The use of teacher guides in education is an example of a service that has been standardized using centrally developed instructions; and our findings demonstrate that the design of these communicated instructions matters for efficiency, a result relevant for other sectors with standardized work practices. Standardized operating procedures and work instructions have mostly concerned the tradeable goods sector, but new information technologies have expanded the span of control of centralized managers in the service sectors (Hsieh and Rossi-Hansberg, 2023). These new technologies have allowed for standardization of services across many establishments by allowing for the dissemination of codified information and best practices (Gawande, 2012). Scripts, checklists, and manuals are also used by medical professionals (Choi et al., 2016; Haynes et al., 2009), airline pilots, and therapists (Eifert et al., 1997).

Our results connect to a large literature examining the impact of information technology on economic productivity (Solow, 1987; Autor et al., 2003; Acemoglu et al., 2014). Advances in new information technologies have reduced barriers to communication in many cases. Interpreted within seminal models of the role of information in firms such as Radner and Zandt (1992) or Bolton and Dewatripont (1994), which view workers as information transmitters within a network, this reduction should always translate directly into greater productivity. However, as has been frequently noted regarding advances in computing technology, productivity gains often lag the adoption of the technology (Solow, 1987; Krishnan et al., 2018). Our results show that the direct impact of technologies that lower communication frictions may not be realized without changes in standards or rules to limit the tendency of workers to share more information than necessary.

Our results also provide the first experimental evidence on the importance of communication skills in a large, multinational organization, adding to a growing literature examining the role of

[2]Structured pedagogy programs are increasingly being adopted based on evidence of their effectiveness and have been identified by the Global Education Evidence Advisory Panel (GEEAP) (Akyeampong et al., 2023) as a "great buy".

social skills in the labor market (Deming, 2017). Our results suggest that when many workers rely on the same, centrally-developed instructions, small editorial investments can result in large impacts. The results beg the question of whether editorial efforts may be undersupplied when their output is reused many times. However, our study is unable to investigate whether lesson designers are incentivized to supply the optimal level of editorial effort.

The remainder of the paper is organized as follows: Section 2 describes the context for the study and its design; Section 3 explains the empirical framework used to analyze the data. Section 4 presents the results of the analysis; and Section 6 concludes.

2 Background and Study Design

2.1 Context

NewGlobe is an education service provider based in Nairobi, Kenya. At the time of the study, the firm operated private schools in Kenya, Nigeria, Uganda, and India, and provided services supporting public schools in Liberia, Nigeria, and Rwanda. Today, they service tens of thousands of teachers, who in turn provide instruction to over two million students.[3]

Schools subscribed to NewGlobe's model use detailed teacher guides in their instruction. Teacher guides are created centrally by staff in Cambridge, Massachusetts, and in the local country of operation. The materials are then disseminated via the internet, and accessed by teachers using tablet e-readers.

Both experiments took place in NewGlobe's Kenyan private schools, operated by their subsidiary, Bridge Kenya (henceforth "Bridge"). Adherence to the teacher guide is an important component of Bridge Kenya's model. Teachers receive specific training in the use of the tablet e-reader. Academy managers (head teachers/principals) conduct two daily observations of each classroom to ensure that teachers are present and using the tablet appropriately. Classroom lesson completion data are also monitored using the tablet.[4] According to NewGlobe, lesson completion data is used primarily for feedback and professional development. However, teachers are required

[3]Since this study, NewGlobe has spun-off management of the fee-charging private schools in Kenya, Nigeria, and Uganda into a separate non-profit organization.

[4]Lesson completion is just one of several metrics used to evaluate teacher performance. In addition, academy managers regularly observe classrooms to gather information on the use of effective teaching strategies, student engagement, and overall classroom management.

to follow the instructions in the guides, and refusal to do so would be grounds for dismissal.

Teacher guides cover every aspect of classroom instruction, including the introduction of new topics, practice problems, and directing students through individual or group work. The guides include prompts for teachers to look around the room and ensure that everyone is paying attention; to have students do work at the front of the class; and to provide positive reinforcement for pupils, such as having the classroom cheer for one another. Lessons are designed to follow the national curriculum. Examples of the lesson guides used in this study can be found in Figures 1 and 2.

Some portions of instruction are not scripted. Teacher guides include modules for individual work where teachers have more discretion over how to engage with the classroom. For example, guides include modules to "Check, Respond, and Leave," where students complete independent work and the teacher checks the work of some students, providing feedback if necessary.[5] In many cases, the amount of time spent on independent work is scripted.

When the study began, Bridge operated approximately 300 private schools in Kenya, serving three years of pre-primary and all eight primary school grades. Bridge schools (i.e., "academies") are distributed across Kenya in 45 out of 47 Kenyan counties.[6]

The population of students in Bridge schools tends to be lower income. At the time of the study, fees were approximately US$100 per year. While the present study has limited data on household wealth or income, Gray-Lobe et al. (2022) report survey results from households with children enrolled at Bridge. These results show that approximately 55 percent of Bridge students have homes with a dirt floor and only 33 percent have electricity at home. The most common occupations reported by caregivers are casual labor (34 percent), small business owner (19 percent), and agriculture (14 percent). Bridge schools tend to be located near lower-income populations. Their schools are split between major cities and smaller towns. In urban and peri-urban centers, Bridge schools are located in informal settlements where lower-income populations reside.

There is considerable variation in class size within Bridge schools. Appendix Figure A1 graphs the distribution of pupil-teacher ratios across classrooms. In the first experiment, the mean class size is 22.5 students per teacher, the median is 21.0, and the standard deviation is 9.5. In Experiment

[5]Teachers are always provided with the correct answers, and training provides teachers with information on how to provide feedback.

[6]During the study, all schools had a single classroom – led by a single assigned teacher – in the grades where the less-detailed guides study was implemented.

2, the mean class size is 24.1, the median is 23.0, and the standard deviation is 8.5. Tables 2 and 3 show that class size is comparable between status quo and less-detailed guide arms in both experiments.

2.2 Experimental Design

Experiment 1. The first experiment altered guides for math instruction for a cohort of lower-primary students as they moved from Grade 2 through Grade 3. The intervention was motivated by feedback from teachers that the guides provided for grade 3 mathematics instruction were overly detailed, by the existing literature showing the potential value of reducing the level of scripting in guides (Piper et al., 2018), and by NewGlobe managers' observations that teaching with very detailed guides posed challenges.

Bridge identified 194 schools for the experiment.[7][8] The analysis removes 10 schools that did not have at least one math score in the two academic terms immediately before the program started.[9]

Out of the 184 schools meeting the inclusion criteria, 89 received shorter guides (treatment), while 95 received "status quo" guides (control). Randomization was stratified by the urbanicity of the school location (rural, urban, or peri-urban), average pupil-teacher ratio, baseline KCPE pass rates, and average lesson completion rates.

The intervention began in September 2018, the final of three academic terms in the 2018 school year, and continued until the second term of the 2019 school year, ending in July 2019 (Figure 3a).

Experiment 2. The second experiment was designed and implemented by NewGlobe after the research team had analyzed the distributional and sub-group tests described in Section 4 below. Specifically, the authors were aware of the interaction with class size, the aggregate compression of test scores, and the compression of between class effects before the second experiment began.

[7] Two hundred schools were originally included, but the stratification procedure described below resulted in 6 small cell strata where treatment assignment was effectively non-random. These are excluded from all analysis.

[8] Some schools were excluded from the experiment for technical reasons that would have interfered with NewGlobe's ability to digitally push content as part of the randomized trial.

[9] Constraining the sample in this way makes use only of data that were available before the randomization, so this restriction does not compromise the integrity of identification from the randomized assignment. We find this characteristic is not predictive of school treatment status. This criterion is intended to restrict the sample to a sub-set of students who are more likely to be actively enrolled through the course of the study. Missing test scores in some cases will indicate that a student has exited Bridge and enrollment records have not been updated. Applying this restriction increases the follow-up rate by approximately 4 percentage points.

The second experiment therefore provided an out-of-sample test of the robustness of the impact observed in the first.

At the time of the second intervention, Bridge operated 112 schools in Kenya, all of which were included in the experiment. Schools were stratified pairwise according to the school's average test score across subjects. Half of the schools were assigned to receive less-detailed guides (treatment) while the other half received the more-detailed, "status quo" guides (control). The intervention started in the second academic term of 2021 and continued through the end of the year (Figure 3b).

2.3 Intervention

Experiment 1. In treated academies, Bridge reduced the daily teacher guide length for grade 3 mathematics instruction. This instruction is separated into two components: 1) "Revision," which is focused on foundational numeracy skills, and 2) "Activities," which presents material from the Kenyan syllabus. On average, mathematics revision daily teacher guides were shortened from 1,500 words to 500 words, and mathematics activities were shortened from three pages to two pages.[10]

The intervention constrained lesson designers to communicate the same material in fewer words, resulting in less detailed guides. Examples of the more- and less-detailed guides can be found in Figure 1. The less-detailed guide used six steps to cover the same general topics that the status quo, more-detailed guide did in fourteen. Importantly, the less-detailed guide also provided only two scripted sentences, compared to nine in the comparison guide.

Less detailed guides gave teachers more opportunities to express concepts in their own words. In the example, teachers were supposed to introduce students to the number 70, introduce its "word form" ("seventy"), and guide the class through the number's relationship to 71, 69, 80, and 60. The status quo, more detailed guide instructed teachers to say "We will write the word form of the number. This is the number seventy... What is 1 more than 70?... What is 1 less than 70?...", along with detailed stage directions indicating when to gesture toward the board and call on students. Meanwhile, the less detailed guides simply state that the teacher should "Discuss the word form of 70 (seventy)..." and "Discuss numbers that are 1 more (71), 1 less (69), 10 more (80), and 10 less (60) with pupils. The guides assume that the teacher would know what a *word form*

[10]These metrics are drawn from internal documents from NewGlobe that describe the interventions.

is, perhaps relying on previous lessons that would have covered this topic. The guide leaves unsaid what form the discussion should take, and whether to call on students.

Most cuts were made to material concerning teacher demonstrations of how to solve problems. Instead of giving detailed information about how to solve the problem, the less-detailed guides provided the problem and its solution, along with an instruction to demonstrate the solution. In Appendix Figure A2, a worked two-digit subtraction problem was reduced from 28 steps to only two. While the instructions do not provide the individual steps to take in solving the problem, the solution strategy – the use of a visual representation of tens and ones – is at least implied by showing the visual representation of the problem. Teachers would have been familiar with this strategy based on its use in previous lessons. Therefore, while the guide left teachers some discretion in how they worked through a problem like this, it may not have been the case that they were left purely to their own devices.

A substantial portion of the guides was not edited at all, and many of the edits removed material that would be unnecessary to explain to someone with a secondary school degree and specialized training. In guides shared with the research team, the teachers' instructions guiding the class through independent practice were unchanged. In many cases, the shorter guides removed instructions that would be unnecessary for most adults with a secondary school degree. For example, in an exercise where the class counted by 5's to the numbers from 5 to 150, the more detailed guides wrote out each number in the sequence, whereas the less detailed guides wrote out the first three numbers in the sequence, and then conveyed the remaining elements using an ellipses.

The more detailed guides may also have contained more errors. At line 9 in the more detailed guide in Figure 1, the guide indicates that 70 minus 10 would be 50. (Line 10 provides the correct answer.) While this suggests that there may be important differences between the guides in terms of the reliability of the information contained within them, it may also illustrate a potential adverse consequence of more detailed instructions: with more details comes more opportunities for error, and with more text, proof-reading effort required to eliminate errors increases.

Experiment 2. In treated classrooms, Bridge reduced the daily teacher guide length for Grade 5 English language instruction. In this experiment, Bridge applied six "simplicity rules" to guide revisions: 1.) keep it linear (avoid any situation where the teacher needs to scroll through the

guide backward), 2.) use lean language, 3.) use consistent language, 4.) do not provide multiple options, 5.) do not provide multi-step instructions, 6.) and use teacher training language properly (See Appendix Table A1 for details).

Though there was no explicit target length for the revised guides, the resulting guides were significantly shorter, similar to the first experiment. In the status quo, the average number of words per guide shrank from 973 to 470, a reduction of 48 percent. The word count of the status quo guides would put it at the 64th percentile of structured pedagogy interventions included in Piper et al. (2018). The intervention moved the guides to approximately the 52nd percentile. The average guide had 118 sentences, compared to 63 in the less-detailed guides. The intervention reduced the average number of words per sentence from 8.34 to 7.46.

The intervention did not affect the readability of the writing, according to Flesch reading ease measure.[11] In both the more- and less-detailed guides, the reading ease score is approximately 80, a level that is considered readable by a US sixth grader.

Less detailed guides (treatment) featured fewer sentences for the teacher to read aloud, as well as reduced teacher instructions.[12] For example, the less-detailed guides removed reminders to complete tasks that are repeatedly performed in every class session (i.e., "procedural clutter").[13] In comparison, the more-detailed guides gave explicit instructions regarding the amount of time teachers were to allocate for individual tasks.

The guidelines used to edit the guides aimed to increase the use of boxed answer keys to provide an easy-to-use reference for information. In the "status quo" materials, answer keys were often embedded within the text. This experiment increased the likelihood that a lesson had a boxed answer key, from 58 to 84 percent. Providing formatted answer keys was intended to make the answers easier for teachers to read and understand.

The guides were also edited to avoid redundancy with the textbooks. Appendix Figure A3 illustrates an example. The status quo version of the guides provided a script to teachers to

[11] Flesch reading ease is probably the most widely used measure of reading difficulty. Scores are obtained from the formula $206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$. The closely related Flesch-Kincaid Grade Level index was developed under a contract with the United States Navy and was originally used to assess the difficulty of technical manuals used by the United States Army and Navy (Kincaid et al., 1975), to "help the understanding of Navy training material when this material must be read by enlisted personnel".

[12] Interestingly, despite reducing the number of read-aloud *sentences*, the fraction of read-aloud *words* per script increased slightly from 40 percent to 42 percent.

[13] Figure 2 shows an example. The original guide started by instructing teachers to "prepare the standard blackboard", which is repeated at the beginning of virtually all lessons.

read aloud that duplicated material that was printed in the textbook, whereas in the less-detailed version, teachers would inform the class that they were going to read from the textbook, and then read from the textbook. In both cases, teachers would have previously instructed the class to open the textbook to the relevant page.

A limitation of the study is that we do not observe how edits to the guides impacted classroom instruction. We do not observe whether teachers understood the instructions better as a result of the edits described above (perhaps because they were easier to follow). Additionally, we do not know whether the removal of information that NewGlobe viewed as superfluous was recalled by the teacher in the moment. Removing details that appear superfluous (for example, because they give instructions that are also given in training) to one person can unintentionally remove a useful reminder (e.g. Haynes et al., 2009; Heffetz et al., 2022). Even if a teacher remembers to perform an unscripted task, they may feel less pressure to execute the task if it is not explicitly listed. It should be noted that the monitoring environment for teachers was unchanged.

2.4 Data

Test score outcomes. For Experiment 1, Bridge administered special assessments, tailored to the specific material covered during the intervention. Bridge developed two tests (posttests) for each academic term of the study – one corresponding to content from supplemental numeracy instruction, and the other to mathematical activities.[14] was administered at the end of each term covering material from the previous Tests were graded and transmitted by another teacher in the school. We focus on posttest assessments in this study. However, we also find evidence of effects on pretest performance (see Appendix Table A4).[15][16]

For Experiment 2 post-intervention test scores came from regular midterm and endterm tests, administered in each academic term, for all subjects. Tests are administered in class, are low-stakes,

[14]Bridge also administered a pretest that mirrored the content of the posttest using slightly different questions.

[15]Interpretation of these effects is challenging. Gains on pretests measuring the ability to transfer knowledge to novel mathematical problems have sometimes been attributed to improvements in higher-order, conceptual skills (Rittle-Johnson et al., 2001). However, the effects on pretests are observed for those corresponding to the students' Revision sessions in the final term of the study. Because Revision sessions review material introduced in Activities sessions, the gains on the Revision pretests may reflect the same gains observed on the Activities posttests.

[16]NewGlobe also shared math midterm and endterm results for Experiment 1. We estimate positive but statistically insignificant effects on these scores (see Appendix Table A4). We prioritize analysis of the posttests because these were specially designed for the study and were the primary outcome according to internal evaluation design documents developed by NewGlobe.

and are primarily used to track pupil performance.[17] Teachers record the results in the tablet and transmit the data to a central office. Scores are communicated to parents.

For this paper, Bridge shared results from English language test scores and related language skill tests, including composition and literacy. We focus here on English language scores since the intervention primarily affected English instruction. We do not find evidence of effects on composition and literacy.

Teacher Behavior Teacher attendance is monitored centrally. We use data shared by NewGlobe on the fraction of classes that a teacher was at school teaching by term. For additional details on this data see Appendix A.

Data on teacher lesson completion comes from telemetric monitoring of teacher progress through digital guides. For the first experiment, NewGlobe shared data on the fraction of lessons that the teacher completed by each teacher. Bridge shared two separate completion rates, one for the math revision class and the other for math activities. We combine these into a single average completion rate for the analysis. For the second experiment, Bridge shared lesson-level data on completion rates and the amount of time per lesson. We use these data to construct a comparable average completion rate at the teacher-term level. We also construct a measure of the time the teacher was in class teaching.

2.5 Sample description, attrition, and balance

In both experiments, treatment and comparison students are equally likely to be observed in the post-intervention tests. In Experiment 1, 59 percent of treatment and comparison students have endline test scores (Panel A of Table 2). Experiment 1 also experienced notable attrition, driven largely by students leaving Bridge at the end of the academic year. The follow-up rate falls from 72 to 57 percent in the comparison group between the 2018 and 2019 academic years. In Experiment 2, 81 percent of comparison and 78 percent of treatment students have endline test scores.

In both experiments, characteristics of treatment and control classrooms are balanced on observed baseline covariates (Table 2 for Experiment 1 and Table 3 for Experiment 2), including baseline test scores. Due to incomplete follow-up, we test for balance conditional on follow-up. We

[17] Individual scores were shared with parents.

also find that treatment and comparison groups are balanced unconditionally.

In Experiment 1, 48 percent of the "status quo" (control) Bridge students were enrolled in schools classified as being in urban or peri-urban locations, compared to 56 percent of treatment students (Table 2). In Experiment 2, 76 percent of control students were enrolled at urban/peri-urban schools, compared to 58 percent of treatment students (Table 2).

2.6 Contemporaneous evaluations

NewGlobe conducted multiple experiments at the same time as the less-detailed guides studies. In the first experiment, grade 8 classrooms also received an intervention that replaced 40-minute periods of classroom review with guided practice sessions for Kenya's primary-school-leaving exam, and grade 4 classrooms received an intervention that assigned students problem sets based on their initial skills instead of giving the whole class the same problem set. In the second experiment, grade 3 classrooms received an intervention that increased the amount of temporally distributed practice in math instruction and grade 2 classrooms received an intervention that increased the amount of repetition in reading assignments for literacy instruction. We assume throughout that these interventions did not influence students in the less-detailed guides experiment. It is unlikely that the introduction of these programs influenced the students and teachers in the less-detailed guides studies, although we cannot empirically test this hypothesis that contemporaneous interventions in other grades and subjects influenced the outcomes studied in the present study. Cross-grade-and-subject spillovers of minor pedagogical changes are rarely tested for by researchers.

3 Empirical framework

We want to estimate the impact of assigning classroom j to shorter guides developed under concision rules, $Z_j = 1$, versus those without, $Z_j = 0$ on test scores Y_i of student i. We estimate the following linear model with potentially heterogeneous effects of less detailed guides

$$Y_i = \alpha + \beta_j Z_j + \varepsilon_{ij} \ . \tag{1}$$

The average effect of concision rules, $E[\beta_j]$, is estimated using Ordinary Least Squares (OLS) regression. The estimated effect $\hat{\beta}$ is identified by random assignment of classrooms to the concision rule guides. We also estimate models of heterogeneous effects including interactions with class size, average class test scores, and within-classroom residualized test scores to examine whether these characteristics predict the impact of the program.

To examine how less-detailed guides affect uniformity of instruction tasks, we examine the distribution of classroom effects. Instructions aim to control the behavior of front-line service workers to ensure that tasks are performed similarly across a large organization, potentially spanning many establishments (Hsieh and Rossi-Hansberg, 2023). Less detailed instructions about a task leaves aspects of its performance up to individual workers. In the most basic model of standardization, codifying how tasks will be performed eliminates variation in outcome that can arise from alternative ways of completing tasks and variation in worker skills. A large empirical literature examines variation in teachers' effectiveness in classroom instruction (Chetty et al., 2014). If, for example, less-detailed guides allow teachers more autonomy over sub-tasks for which concision rules remove direct instructions, relying more on teacher discretion effects on the dispersion in aggregate classroom performance may reflect variance in teachers' ability to use this time.

We test for changes in the dispersion of value-added using two measures of classroom effects. The first is the class average test score. Classroom means potentially contain bias on average due to selection effects, but given random assignment to the less-detailed guides condition, the selection bias of this value-added estimate is unlikely to be associated with guide detail. We also estimate observational value-added as the estimated classroom effect from a model with a lagged test score control. A large literature shows that lagged score controls can remove selection bias from school and teacher value-added estimates (Chetty et al., 2014; Deming, 2014; Angrist et al., 2017).

Finally, we also test whether concision rules have heterogeneous impacts on individual students. The level of detail in guides could affect the within-classroom distribution through several channels. Teachers could use saved time to help students who have fallen behind catch up. They could also use their knowledge of students' personalities and knowledge to explain things in ways that are more adapted to the classroom's needs. We examine impacts on the residualized (within class) baseline test scores of students and compare the distribution of the residualized test scores between schools.

4 Results

This section presents results on the average treatment effect of making guides less detailed on student test scores, heterogeneous effects, and effects on dispersion in test scores. We also examine effects on teacher absences and lesson completion.

4.1 Average treatment effect of less-detailed guides on test scores

Less-detailed guides increased subject test scores by 0.17 standard deviations (Table 4 for Experiment 1 and 5 for Experiment 2), with effects growing over time (Panel A of Figure 4). In the first experiment, there was no discernible effect after one term (midline 1) (Figure 4). After two terms, the estimate reached 0.10 standard deviations, although this effect is not statistically significant. In the second experiment, the effect seems to have emerged more quickly (Panel B of Figure 4). The estimated effect after one term (midline 1) was only 0.09 and not statistically significant. After one and a half terms (midline 2) the effect was 0.15 standard deviations.

In both cases, the impacts do not emerge as statistically significant until the final period for which data is available, although the estimated effects across periods within each experiment are statistically indistinguishable. This pattern of effects may indicate that teachers required time to adjust to the new approach.

In any case, it is noteworthy that the impact in the second experiment emerges more rapidly than that in the first experiment. This result is consistent with the view that the approach to editing the guides in the second experiment, which reflects learning from the first experiment, may have been more effective, and that the final term effect from the first experiment could reflect a lower bound of the longer-run effect of shorter guides.

4.2 Heterogeneous effects

The effect of less-detailed guides on endline test scores is largest in classrooms with larger class sizes. In the first experiment, the effect is approximately 0.016 standard deviations larger for each student in the classroom (Table 4). The effect for a student at the median of the class size distribution (18) is 0.138 standard deviations. In a class with 10 more students, the effect of less-detailed guides is 0.160 standard deviations. In the second experiment, the effect is 0.03 standard deviations larger

for each additional student Table 5. The effect of less-detailed guides for a student at the median of the class size distribution (24) is 0.327 standard deviations. This effect is almost twice as large in classrooms with 30 versus 20 students.

To examine heterogeneity in terms of baseline test scores, we distinguish between classroom and within-classroom differences to mirror the discussion of the distributional effects below. Specifically, we test separately for heterogeneity in the effect depending on the mean test score and depending on a student's residualized (subtracting out the classroom effect) test score.

We find little evidence that the effects of less-detailed guides within classrooms vary depending on students' residualized test scores. In the first experiment, we estimate the effect on a student at the classroom mean to be 0.189 standard deviations. The estimate of the interaction term indicates that the effect on a student one standard deviation above the mean is 0.002 standard deviations greater. In the second experiment, the effect on a student at the classroom mean is 0.175 standard deviations. The effect on a student one standard deviation above the mean is 0.059 standard deviations greater. In neither case is the interaction effect statistically distinguishable from zero.

Interactions with classroom means are large in magnitude although not statistically significant. In the first experiment, the estimated interaction effect with classroom mean scores is -0.206. In the second experiment, the estimate is 0.423 standard deviations. While we cannot reject the null hypothesis that classroom means are unrelated to the effect of less-detailed guides, it is noteworthy that the sign of the results align with the distributional effects reported below.

4.3 Effects on the dispersion of value-added

In the first experiment, less-detailed guides compressed the distribution of classroom effects using classroom means, reducing mass in the left tail of the distribution (Panel A of Figure 5). The status quo standard deviation is equivalent to 0.60 standard deviations of the student test score distribution. In the less-detailed guide condition, this falls to 0.43. An F-test of the equality of the standard deviations rejects the null hypothesis at less than the one percent level. Levene's test (which is more robust to non-normality) yields a *p-value* of 0.02. Visual comparison suggests the compression is due to less mass in the left tail. Combined with the heterogeneity estimates reported above, these results suggest that less-detailed guides compressed the value-added distribution by reducing the prevalence of lower-performing classrooms. We find similar results using observational

value-added estimates derived from a model with lagged test scores. Going from more- to less-detailed teacher guides reduces the standard deviation of the classroom value-added estimate from 0.51 to 0.42, a difference that is statistically significant at less than the 10 percent level according to the F-test, although Levene's test fails to reject the null hypothesis.

In the second experiment, less-detailed guides are associated with *more* dispersed classroom effects (Panel B). The status quo standard deviation of class effects using classroom means is equivalent to 0.43 standard deviations of the student test score distribution. In the less-detailed guide arm, it is equivalent to 0.55 standard deviations. While there are no schools in the comparison group with classroom means above one standard deviation, there are five in the less-detailed guide condition. An F-test rejects the null hypothesis that the effects are equal at less than the 10 percent level, although Levene's test does not.[18] Although we cannot confidently reject the hypothesis that less-detailed guides had no impact on the dispersion in classroom means, the larger impacts on the bottom of the distribution from the first experiment were not replicated in the second. Visually, the less-detailed guides may have increased mass in the right tail. Together with the large and marginally significant interaction effect with mean classroom baseline scores, it seems less-detailed guides may have increased the dispersion of classroom effects with some classrooms succeeding much more than others. Again, observational classroom value-added estimates obtained after controlling for lagged test scores reveal a similar pattern. Going from more- to less-detailed teacher guides increased the standard deviation of estimated classroom value-added from 0.31 to 0.40. The difference is statistically significant at less than the 10 percent level using the F-test, but not statistically significant when using Levene's test (although the *p-value* is 0.11).

In both cases, the change in the distribution due to less-detailed guides is observed in both small and large classes. Appendix Tables 6 and 7 report comparable results for classes that are above and below the median class size level. In the first experiment, the standard deviation of class effects fell from 0.63 to 0.43 in classrooms with above-median class sizes, and fell from 0.55 to 0.44 in below-median classrooms. In the second experiment, the standard deviation of class effects increased from 0.41 to 0.52 in classrooms with above-median class sizes and from 0.44 to 0.60 in those with below-median class sizes.

[18] A Kolmogorov-Smirnov test that the distributions are equal fails to reject the null hypothesis (unreported).

4.4 Effects on the within-classroom distribution

In both experiments, the dispersion in test scores within classrooms fell. In the first experiment, the standard deviation of residualized test scores is 0.82 in status quo classrooms, compared to 0.77 (Panel A of Figure 5). The F-test *p-value* is 0.07, while that from Levene's test is 0.15. In the second experiment, the standard deviation in the status quo was 0.84, compared to 0.79 in the less-detailed guide arm. The F-test *p-value* is 0.1 and that from Levene's test is 0.21.

The compression within classrooms is concentrated among high-PTR classrooms. Appendix Tables 6 and 7 show that in both experiments, the within-classroom distribution fell more in those classrooms with above median PTR.

4.5 Effects on teacher attendance

In the first experiment, there is no evidence that teacher attendance is affected by the introduction of less-detailed guides. In Term 1 of 2019 (the second term of the study), teachers in the status quo condition were present and teaching 82 percent of the time (Figure 6). Less detailed guides are associated with approximately a 0.9 percentage point reduction, although this estimate is not statistically distinguishable from zero. In term 2, 2019 (the final term of the study) teachers in the status quo condition were in class teaching 91 percent of the time, and less-detailed guides are associated with approximately a 0.1 percentage point increase (also not statistically significant).[19]

In the second experiment, teacher attendance appears to have increased in response to the less-detailed guide condition, although the effect is only statistically significant for the final term. In Term 2 of 2021, teachers in the status quo condition were in class teaching approximately 89 percent of the time, and less-detailed guides are associated with a 2.3 percentage point increase. This effect is not statistically distinguishable from zero. In Term 3, teachers in the status quo condition were in class teaching approximately 90 percent of the time, and less-detailed guides are associated with a 5.8 percentage point increase. This effect is statistically significant at less than the 5 percent level. The impact on teacher attendance translates into an additional four hours of instruction from the class's assigned teacher (Panel B Column 3 Table 8).

[19]Bridge did not share termly attendance data for 2018, so results for the first term of the study are not reported.

4.6 Effects on lesson completion

Lesson completion is linked to the classroom's *assigned* teacher, complicating the analysis. In the first experiment, NewGlobe shared the fraction of lessons completed by the classroom's teacher when that teacher was present and teaching. In the second experiment, we base our analysis on raw lesson completion data where absence of a classroom's teacher results in a value of zero for lesson completion. In neither case do not observe completion of lessons by substitute teachers. This poses two challenges in interpreting the results from this analysis. First, it is not possible to construct a measure of the number of lessons that students received. Second, estimates of the classroom-level lesson completion rate can reflect changes in teacher attendance. To address the latter issue, we report estimates of the effect of a measure of the lesson completion rate conditional on teacher attendance by rescaling the completion rate by the teacher attendance rate. Interpretation of these estimates requires an assumption that teachers' absences from class are independent of the likelihood of a lesson's completion.

We focus the analysis on lesson completion in the final term of the study because a malfunction in the recording of lesson completion led to many lessons being recorded as incomplete in Term 2 of 2021. For symmetry, we also focus on results from the final term of Experiment 1 (the results are similar when including both terms for which lesson completion data are available in Experiment 1). Another reason for focusing on the final term is that the impacts on lesson completion may not be seen until teachers have a chance to adapt to a new convention.

In the first experiment, lesson completion is lower in the less-detailed guide group than in the status quo, although this difference is not statistically significant at a conventional level. While approximately 66.7 percent of lessons were completed in the status quo group, less-detailed guides are associated with 3.9 percentage points lower completion rates (Panel A Column 5 of Table 8).

In the second experiment, lesson completion is higher in the less-detailed guide group. Again approximately 66.7 percent of lessons were completed in the status quo condition, and less-detailed guide teachers completed 8.4 percentage points more lessons (Panel B Column 5). This effect is statistically significant at less than the five percent level. However, the effect may reflect the increase in teacher attendance noted above. The effect on lesson completion conditional on teachers' attendance is 6.1 percentage points. This effect, however, is not statistically significant at a conventional

level.

We find no evidence that the effects on teacher attendance or lesson completion varies with class size (even columns in Table 8), suggesting that these teacher behaviors are unlikely to be mediators of the effect. Bridge employs substitute teachers when assigned teachers are absent, so it is unclear if the teacher attendance effect would have impacted instruction received by students.

5 A model of task standardization

In a naive model of codification of teaching tasks, giving detailed instructions for how to perform a task should strictly reduce dispersion in how it is performed. Imagine teachers are all left to their own devices to complete a task. That task can be performed many ways and so there's a distribution of the quality of its completion μ_j (measured in terms of the student learning gains from completion of the task). Variation in how teachers might complete the task without explicit guidance yields a distribution $F(\mu)$, which is assumed to be distributed uniformly on the interval $[0, \gamma]$.

A planner can intervene to provide explicit instructions codifying that the task be completed in a particular way. Assuming the instructions are adhered to, the resulting level of quality is $\alpha > 0$.

Lemma 1 *If the planner aims to maximize average learning gains, then the planner should provide explicit instructions whenever* $\alpha > \frac{\gamma}{2}$.

With instructions, the average learning gain is α, whereas without instructions, the average learning gain is $E[\mu] = \frac{\gamma}{2}$.

Interpreted within this model, the negative effect observed from both experiments indicates that (on average) for those tasks that saw explicit details removed, the quality of the task as instructed was less than the expected quality of what teachers could do without instructions $\alpha < \frac{\gamma}{2}$.

By construction, the negative average effect should be accompanied by a reduction in the dispersion in outcomes and, notably due in part by a reduction in the number of teachers completing the task at qualities below α. The model is therefore mostly consistent with the results from the first experiment, where the increase in classroom effects comes from an increase in the number of high-performing classrooms.

However, the effect of the intervention on the dispersion of test scores in the first experiment directly contradicts the model. Providing less-detailed guides led to a more compressed distribution of classroom effects and the compression was driven by a reduction in very low-performing classrooms. Visually at least, the same reduction in very low-performing classrooms can also be observed in the second experiment.

The model can accommodate these results by noting that teachers may need to spend cognitive effort and time reading instructions, which can crowd out other important activities. Teachers are assumed to perform a complementary task (such as perspective-taking, classroom management, providing unscripted and potentially individualized feedback) $x \in \{0, 1\}$. The variation in classroom effects arises from heterogeneous costs of reading instructions. A fraction p of teachers experience a depletion of cognitive bandwidth and execute the task at a quality $\delta\alpha$. Now the quality of the task when explicit instructions are provided is $q_j = p\delta\alpha + (1-p)\alpha$, and $q_j = \mu_j$ without.

Lemma 2 *The planner now prefers to provide explicit instructions only if* $p\delta\alpha + (1-p)\alpha > \frac{\gamma}{2}$.

In the augmented model, each teacher can be characterized by a vector of "skills" (δ_j, μ_j).

Lemma 3 *There may exist some teachers for which* $\alpha\delta_j - \mu_j < 0$.

In other words, there may be some mass of teachers for whom the provision of instructions is strictly harmful. For example, imagine that δ_j is a function of μ_j. This could arise if teachers for whom the uninstructed quality is higher are more prepared (e.g., have reviewed the instructions beforehand) or have more prior knowledge (e.g., it is easier to skim text if one is more familiar with the topic). For example, if $\delta_j = \frac{\mu_j}{\gamma}$, then as long as $\alpha < \gamma$, instructions will induce a distribution of $q_j = \delta_j\alpha$, $G(q_j)$, that is first-order stochastically dominated by $H(q_j)$, the distribution without codified instructions.

This extension of the model may also provide insight into the larger effects observed in larger classrooms. Teachers in large classrooms may be more constrained both in terms of cognitive bandwidth and time. Teachers may need to spend time on classroom management to address disruptive behaviors. Some tasks, such as providing individualized feedback to students, may require more time in large classrooms.

6 Conclusion

Editing teacher guides to make them shorter and less detailed produced moderate and robust positive impacts on student academic skills. Gains were observed in both Grade 3 mathematics and Grade 5 English instruction, suggesting that these impacts may apply broadly across academic subjects. In both studies, the effect is concentrated among classrooms with high pupil-teacher ratios. There is also some evidence that shorter guides compressed the within-class distribution of classrooms.

These results are directly relevant to the design of guidelines for teachers, a common component of structured pedagogy programs. Structured pedagogy has been adopted across a variety of contexts (Piper et al., 2018; Stockard et al., 2018), and are rated as "great buys" by the Global Education Evidence Advisory Panel (GEEAP) (Akyeampong et al., 2023). Piper et al. (2018) noted an association between such programs' effects and script length; our study contributes to this literature by providing experimental evidence that less-detailed guides improve learning outcomes.

Scripting less leaves more aspects of instruction to teacher discretion. To the extent that the quality of a teacher's unscripted instruction diverges from that of scripted instruction, the variance in outcomes will increase. While less-detailed guides reduced dispersion in between classroom effects with fewer low-performing schools in the first experiment, they may have increased dispersion in the second experiment with more high-performing schools.

There are several reasons why the two experiments may have had different effects on the distribution of classroom effects. First, the teachers of older grades are specialized in specific subjects, and hiring favors those with more subject knowledge. These teachers may therefore have more to contribute when using their discretion. Having more expertise may mitigate the effects of unclear communication (Pinker, 2014). Second, this difference across the two experiments could reflect differences in the communication codes used. The first experiment applied a space constraint whereas the second experiment applied concision rules such as presenting information in tabular form. If these materials can be leveraged for greater effect by some teachers, this can increase variation in classroom mean scores.

Another hypothesis is that the level of detail in the guides was appropriate when teachers were new, but has become less important with time. Bridge Kenya was established in 2009 and expanded

rapidly from 2013 to 2018. Initially, it hired many teachers with only a secondary school degree based on the founders' view that traditional teacher training was not important as long as new teachers received appropriate training, support, monitoring, and incentives (Rangan and Lee, 2010). As a result, many teachers would have had an unusally low level of prior teaching experience. Over time, however, teachers may have developed skills on the job. Many of Bridge's teachers have also received teacher training certification while on the job, and new hires often have this training when they enter (Gray-Lobe et al., 2022). With more experience and training, teachers in 2018 may have needed less detailed instructions than when Bridge was starting out.

The impacts on teacher attendance in the second experiment suggest that efforts to make instructions less detailed led to improvements in job satisfaction. This effect could reflect a variety of mechanisms. First, teachers may enjoy a greater sense of autonomy (Harju et al., 2021; Adhvaryu et al., 2019) if the guides allow them to improvise more. This interpretation could explain why the effect is observed in the first but not the second study. Fifth-grade teachers are subject teachers and would likely have more specific interests in their subject. Teachers with more skill and interest in a subject may find autonomy to be more of an amenity. Second, teachers may get more satisfaction from their job if they see that they are more effective as instructors so the attendance effect could be an endogenous response to the student test score gains. Third, they may find their job more enjoyable if they don't have to parse long and detailed instructions. The task overload literature has emphasized the potential negative impacts on worker welfare, including higher levels of stress and burnout. Our results suggest that imposing artificial barriers to communication can enhance productivity. While communication costs are often conceptualized as being barriers to the division of labor (Becker and Murphy, 1992) and scale (Bolton and Dewatripont, 1994), technologies that ease the flow of communication can pose new problems due to information overload (Karr-Wisniewski and Lu, 2010).

This study provides experimental evidence on the potential negative impact of information overload. The positive relationship between the effect and class size suggests the effect of less-detailed instructions may arise in part from reduced cognitive load requirements on teachers. Larger class sizes are associated with more disruptive classrooms and behavioral problems to manage (Lazear, 1999), less personal connections to students (Finn et al., 2003), and higher levels of reported teacher stress (French, 1993; Huang et al., 2022). A large literature demonstrates cognitive strain

can impede communication skills including both the abilities to give instructions (Kurinec et al., 2019; Roßnagel, 2000, 2004; Kurinec et al., 2019), follow instructions (Yang et al., 2015), and receive feedback (Roßnagel, 2004). In the setting of this study, inefficient communication from lesson planners to teachers results in worse instruction from teachers to students, suggesting that the negative impacts of poor communication can reverberate. A taxed reader will write less effectively to others.

These results are unlikely to reflect selective attrition for several reasons. First, we find no evidence that the intervention influenced the level of follow-up in student assessments. It is unlikely that the intervention would have affected the decision to change schools entirely, given that it affected only a single subject. Second, we find comparable effects in Experiment 2, which had a relatively high level of follow-up. Third, as noted in Section 2.5, the samples appear to be similar conditional on follow-up.

References

Acemoglu, D., D. Autor, D. Dorn, G. H. Hanson, and B. Price (2014): "Return of the Solow Paradox? IT, Productivity, and Employment in US Manufacturing," American Economic Review, 104, 394–99.

Adhvaryu, A., T. Molina, and A. Nyshadham (2019): "Expectations, Wage Hikes, and Worker Voice: Evidence from a Field Experiment," Working Paper 25866, National Bureau of Economic Research.

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

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," The Quarterly Journal of Economics, 132, pp. 871–919.

Arnold, M., M. Goldschmitt, and T. Rigotti (2023): "Dealing with Information Overload: A Comprehensive Review," Frontiers in Psychology, 14, 1122200.

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The Skills Content of Recent Technological Change: An Empirical Exploration," The Quarterly Journal of Economics, 118, 1279–1333.

Becker, G. S. and K. M. Murphy (1992): "The Division of Labor, Coordination Costs, and Knowledge," The Quarterly Journal of Economics, 107, 1137–1160.

Bold, T., D. Filmer, G. Martin, E. Molina, C. Rockmore, B. Stacy, J. Svensson, and W. Wane (2017): "What do teachers know and do? Does it matter? Evidence from Primary schools in Africa," .

Bolton, P. and M. Dewatripont (1994): "The Firm as a Communication Network," The Quarterly Journal of Economics, 109, 809–839.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," American Economic Review, 104, 2633–79.

Choi, H.-Y., E. H. Kim, J. Yoo, K. Lee, D. Song, young Dae Kim, H.-J. Cho, H. S. Nam, K. Y. Lee, H. S. Lee, and J. H. Heo (2016): "Decision-Making Support Using a Standardized Script and Visual Decision Aid to Reduce Door-to-Needle Time in Stroke," Journal of Stroke, 18, 239–241.

Deming, D. J. (2014): "Using School Choice Lotteries to Test Measures of School Effectiveness," American Economic Review, 104, 406–11.

——— (2017): "The Growing Importance of Social Skills in the Labor Market," Quarterly Journal of Economics, 132, 1593–1640.

Dresser, R. (2012): "The Impact of Scripted Literacy Instruction on Teachers and Students," Issues in Teacher Education.

Eifert, G. H., D. Schulte, M. J. Zvolensky, C. Lejuez, and A. W. Lau (1997): "Manualized behavior therapy: Merits and challenges," Behavior Therapy, 28, 499–509.

Fazzio, I., A. Eble, R. L. Lumsdaine, P. Boone, B. Bouy, P.-T. J. Hsieh, C. Jayanty, S. Johnson, and A. F. Silva (2021): "Large learning gains in pockets of extreme poverty: Experimental evidence from Guinea Bissau," Journal of Public Economics, 199.

Finn, J. D., G. M. Pannozzo, and C. M. Achilles (2003): "The "Why's" of Class Size: Student Behavior in Small Classes," Review of Educational Research, 73, 321–368.

French, N. K. (1993): "Elementary teacher stress and class size," Journal of research and development in education, 26, 66–73.

Gawande, A. (2012): "Big Med," The New Yorker.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. Ozier (2022): "Can education be standardized? Evidence from Kenya," Becker-Friedman Working Paper Series.

Harju, J., S. Jäger, and B. Schoefer (2021): "Voice at Work," Working Paper 28522, National Bureau of Economic Research.

Haynes, A. B., T. G. Weiser, W. R. Berry, S. R. Lipsitz, A.-H. S. Breizat, E. P. Dellinger, T. Herbosa, S. Joseph, P. L. Kibatala, M. C. M. Lapitan, A. F. Merry, K. Moorthy, R. K. Reznick, B. Taylor, and A. A. Gawande (2009): "A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population," New England Journal of Medicine, 360, 491–499, pMID: 19144931.

Heffetz, O., T. O'Donoghue, and H. S. Schneider (2022): "Reminders Work, but for Whom? Evidence from New York City Parking Ticket Recipients," American Economic Journal: Economic Policy, 14, 343–70.

Holmstrom, B. (1980): "On The Theory of Delegation," Discussion Paper 438, Evanston, IL.

Hsieh, C.-T. and E. Rossi-Hansberg (2023): "The Industrial Revolution in Services," Journal of Political Economy Macroeconomics, 1, 3–42.

Huang, Y., E. Richter, T. Kleickmann, and D. Richter (2022): "Class size affects preservice teachers' physiological and psychological stress reactions: An experiment in a virtual reality classroom," Computers & Education, 184, 104503.

Karr-Wisniewski, P. and Y. Lu (2010): "When more is too much: Operationalizing technology overload and exploring its impact on knowledge worker productivity," Computers in Human Behavior, 26, 1061–1072, advancing Educational Research on Computer-supported Collaborative Learning (CSCL) through the use of gStudy CSCL Tools.

Kincaid, J. P., J. Fishburne, Robert P., R. L. Rogers, and B. S. Chissom (1975): "Derivation of New Readability Formulas: (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," Tech. rep., Chief of Naval Technical Training, Naval Air Station Memphis, Millington, Springfield, distributed by NTIS.

Krishnan, M., J. Mischke, and J. Remes (2018): "Is the Solow Paradox back?" McKinsey Quarterly, article.

Kurinec, C. A., A. Wise, C. A. Cavazos, E. M. Reyes, and C. A. Weaver (2019): "Clarity under cognitive constraint: Can a simple directive encourage busy speakers to avoid ambiguity?" Language and Cognition, 11, 621–644.

Lazear, E. P. (1999): "Globalisation and the Market for Team-Mates," The Economic Journal, 109, C15–C40.

Pinker, S. (2014): "Why Academics Stink at Writing," The Chronicle Review.

Piper, B., E. Jepkemei, D. Kwayumba, and K. Kibukho (2015): "Kenya's ICT Policy in Practice: The Effectiveness of Tablets and E-Readers in Improving Student Outcomes," Forum for International Research in Education, 2, 3–18.

Piper, B., Y. Sitabkhan, J. Mejia, and K. Betts (2018): "Effectiveness of Teachers' Guides in the Global South: Scripting, Learning Outcomes, and Classroom Utilization," RTI Press - Occasional Paper.

Radner, R. and T. V. Zandt (1992): "Information Processing in Firms and Returns to Scale," Annales d'Économie et de Statistique, 265–298.

Ragu-Nathan, T. S., M. Tarafdar, B. S. Ragu-Nathan, and Q. Tu (2008): "The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation," Information Systems Research, 19, 417–433.

Rangan, V. K. and K. Lee (2010): "Bridge International Academies: A School in a Box," Harvard Business School Case Study.

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.

Romero, M., J. Sandefur, and W. A. Sandholtz (2019): "Outsourcing Education: Experimental Evidence from Liberia," American Economic Review.

Rossnagel, C. (2000): "Cognitive load and perspective-taking: applying the automatic-controlled distinction to verbal communication," European Journal of Social Psychology, 30, 429–445.

——— (2004): "Lost in thought: cognitive load and the processing of addressees' feedback in verbal communication." Experimental psychology, 51 3, 191–200.

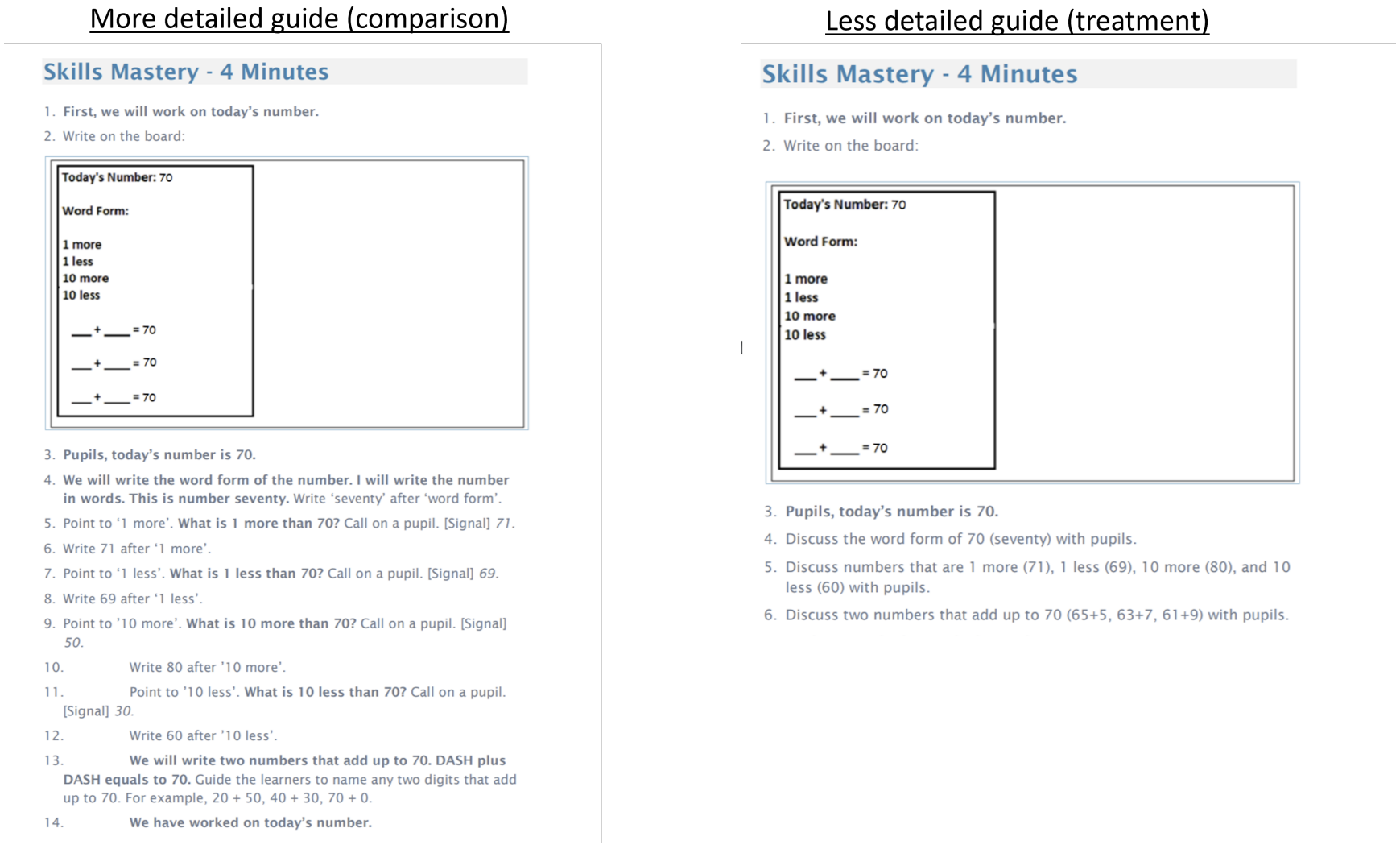
Solow, R. (1987): "We'd Better Watch Out," New York Times Book Review, 36.

Stockard, J., T. W. Wood, C. Coughlin, and C. Khoury (2018): "The Effectiveness of Direct Instruction Curricula: A Meta-Analysis of a Half Century of Research," Review of Educational Research, 88.

Yang, T., R. Allen, and S. Gathercole (2015): "Examining the role of working memory resources in following spoken instructions," Journal of Cognitive Psychology, 28, 1–13.

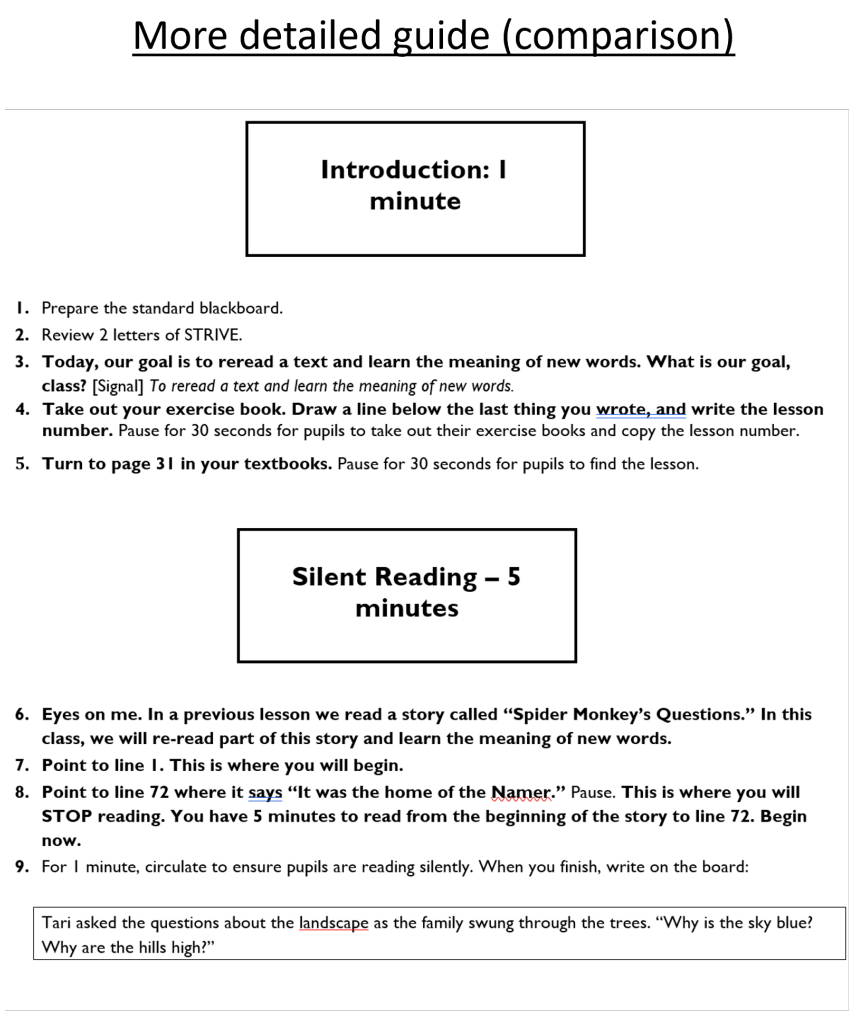
Figures & Tables

Figure 1: Comparison of teacher guides in control Experiment 1



Notes: The more detailed guide was the status quo version. Bolded instructions would be read aloud to the class. Unbolded instructions are to the teacher. This lesson discussed the word 70, giving the "word form" and explaining its arithmetic relationship to nearby numbers. Prior lessons would have introduced teachers and students to the concept of a number's "word form". While the less-detailed guide removes instructions to call on pupils, it should be noted that this type of interaction is present in virtually all instruction and teachers would receive specific training in its use.

Figure 2: Notes: Comparison of teacher guides Experiment 2



More detailed guide (comparison)

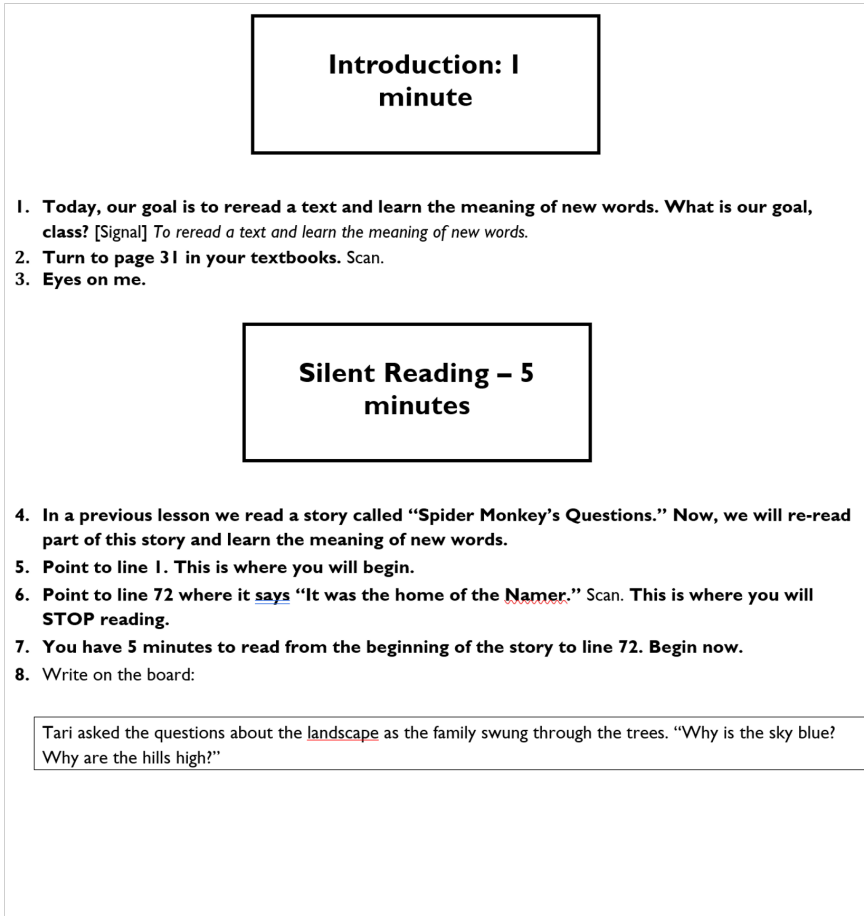
Introduction: 1 minute

1. Prepare the standard blackboard.
2. Review 2 letters of STRIVE.
3. **Today, our goal is to reread a text and learn the meaning of new words. What is our goal, class?** [Signal] *To reread a text and learn the meaning of new words.*
4. **Take out your exercise book. Draw a line below the last thing you wrote, and write the lesson number.** Pause for 30 seconds for pupils to take out their exercise books and copy the lesson number.
5. **Turn to page 31 in your textbooks.** Pause for 30 seconds for pupils to find the lesson.

Silent Reading – 5 minutes

6. **Eyes on me. In a previous lesson we read a story called "Spider Monkey's Questions." In this class, we will re-read part of this story and learn the meaning of new words.**
7. **Point to line 1. This is where you will begin.**
8. **Point to line 72 where it says "It was the home of the Namer."** Pause. **This is where you will STOP reading. You have 5 minutes to read from the beginning of the story to line 72. Begin now.**
9. For 1 minute, circulate to ensure pupils are reading silently. When you finish, write on the board:

Tari asked the questions about the landscape as the family swung through the trees. "Why is the sky blue? Why are the hills high?"



Less detailed guide (treatment)

Introduction: 1 minute

1. **Today, our goal is to reread a text and learn the meaning of new words. What is our goal, class?** [Signal] *To reread a text and learn the meaning of new words.*
2. **Turn to page 31 in your textbooks.** Scan.
3. **Eyes on me.**

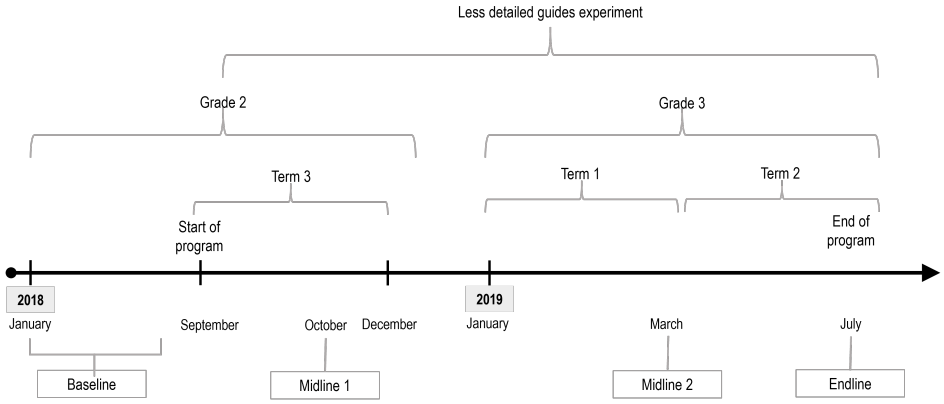
Silent Reading – 5 minutes

4. **In a previous lesson we read a story called "Spider Monkey's Questions." Now, we will re-read part of this story and learn the meaning of new words.**
5. **Point to line 1. This is where you will begin.**
6. **Point to line 72 where it says "It was the home of the Namer."** Scan. **This is where you will STOP reading.**
7. **You have 5 minutes to read from the beginning of the story to line 72. Begin now.**
8. Write on the board:

Tari asked the questions about the landscape as the family swung through the trees. "Why is the sky blue? Why are the hills high?"

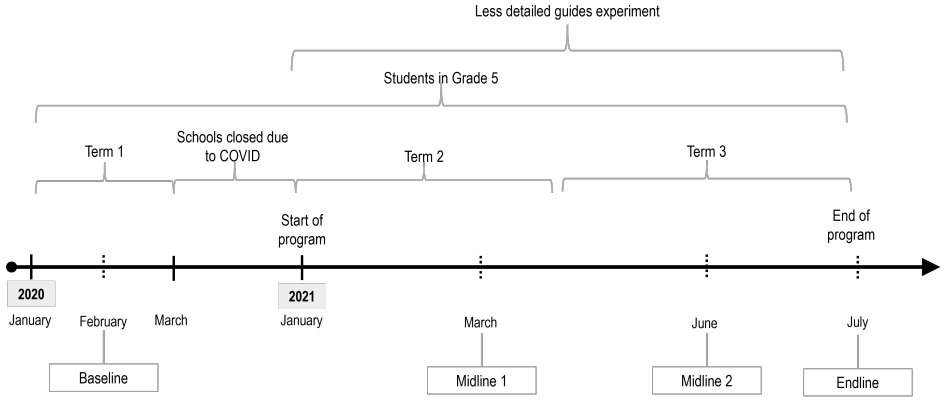
Notes: The first two instructions in the more detailed guide are activities that represent tasks that are always performed at the start of the school day. STRIVE is part of a positive behavioral management strategy. STRIVE stands for Sit tall, Track the teacher with your eyes and body, Respond to questions, Inquire, Visualize success, Ears are for listening. Teachers often begin class by reviewing two of the letters. The standard blackboard lists the acronym STRIVE vertically along the left side of the board and indicates the day of the week, date, and subject on the top of the board. The standard blackboard also lists the current tally of STRIVE points in the bottom right corner of the board. Strive points are awarded to groups of students (those seated in the same row) when they demonstrate appropriate behavior.

Figure 3



(a)

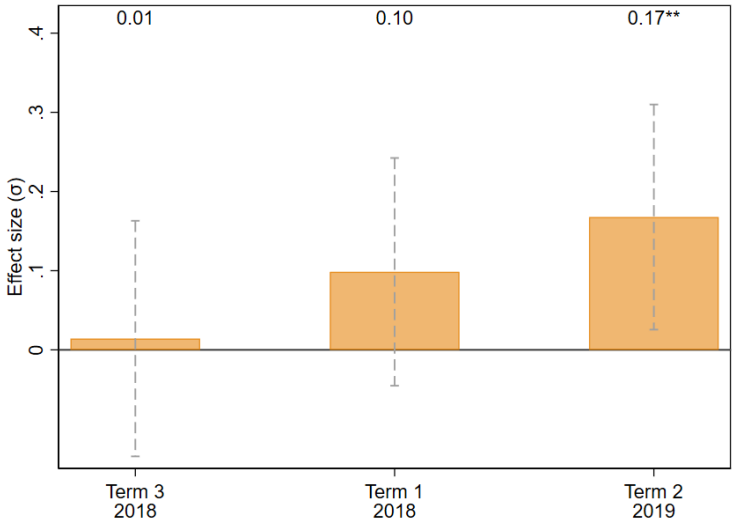
Experiment 1 timeline.



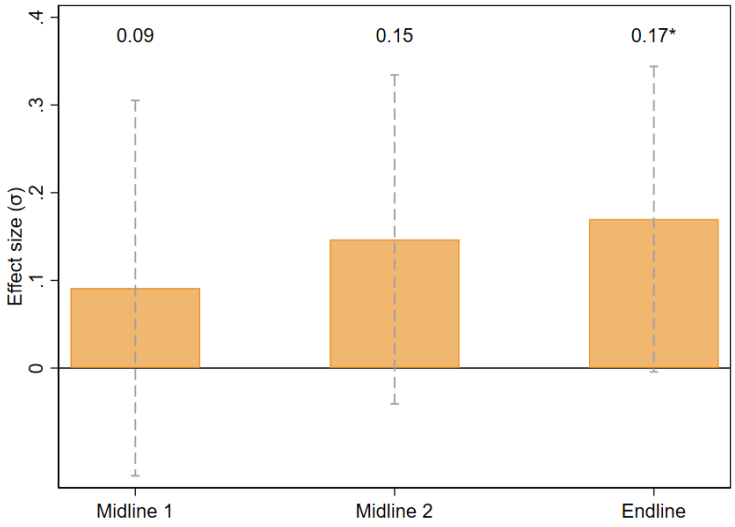
(b) Experiment 2 timeline.

Notes: This figure shows the timeline of the intervention and test score data collection. Experiment 1 spanned two academic years, starting in the final term of the 2018 academic year and ending in the second term of 2019. Experiment 1 began in the second term of the 2020 academic year, which was interrupted in March, 2020 due to the COVID-19 pandemic and continued through the end of the academic year in July. The boxed text below the figure indicates the timing of test score data collection used in the paper.

Figure 4

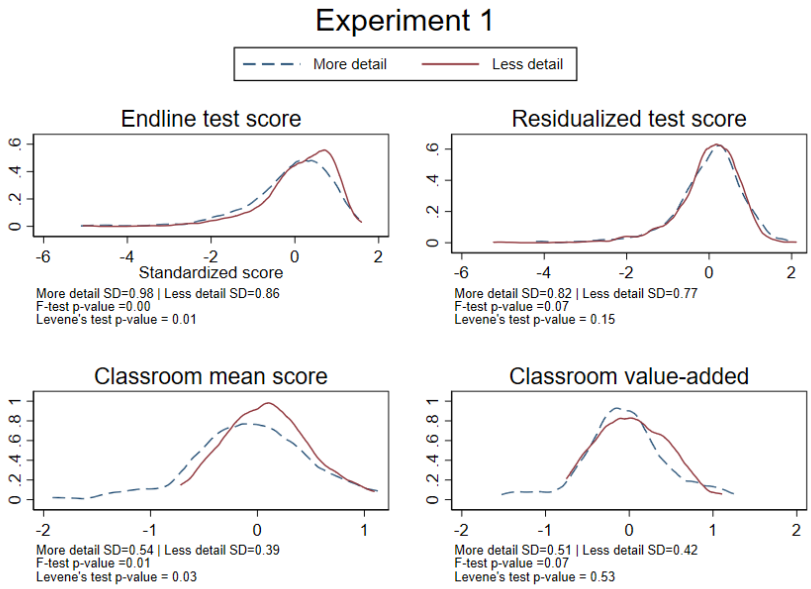


(a) Experiment 1

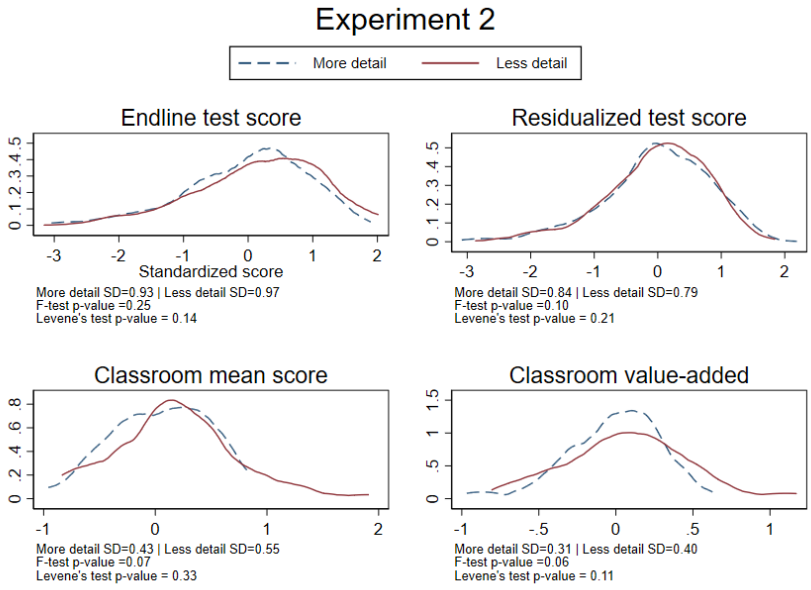


(b) Experiment 2

Notes: Coefficient plot of the effect of less detailed teacher guides on test scores. The top panel reports the effect from the first experiment on math scores. The bottom panel reports the effect from the second experiment on English language scores. Estimates in experiment 1 control for baseline math score. Estimates in Experiment 2 control for baseline language score and randomization strata indicators. The line indicates the 90% confidence interval. Standard errors are clustered at the school level. Point estimates are reported above. *, **, and *** indicate statistical significance at the 10, 5 and 1% level.



(a)



(b)

Figure 5: Kernel density plots of test scores for status quo (more detailed) and less-detailed guide classrooms. The top figure in each panel shows the distribution of test scores. The bottom left figure shows the distribution of classroom mean scores. The bottom right figure shows the distribution of test scores after partialling out the classroom mean. In Experiment 1, the test score is the Term 3 posttest index. In Experiment 2 the test score is the Term 2 English language endterm.

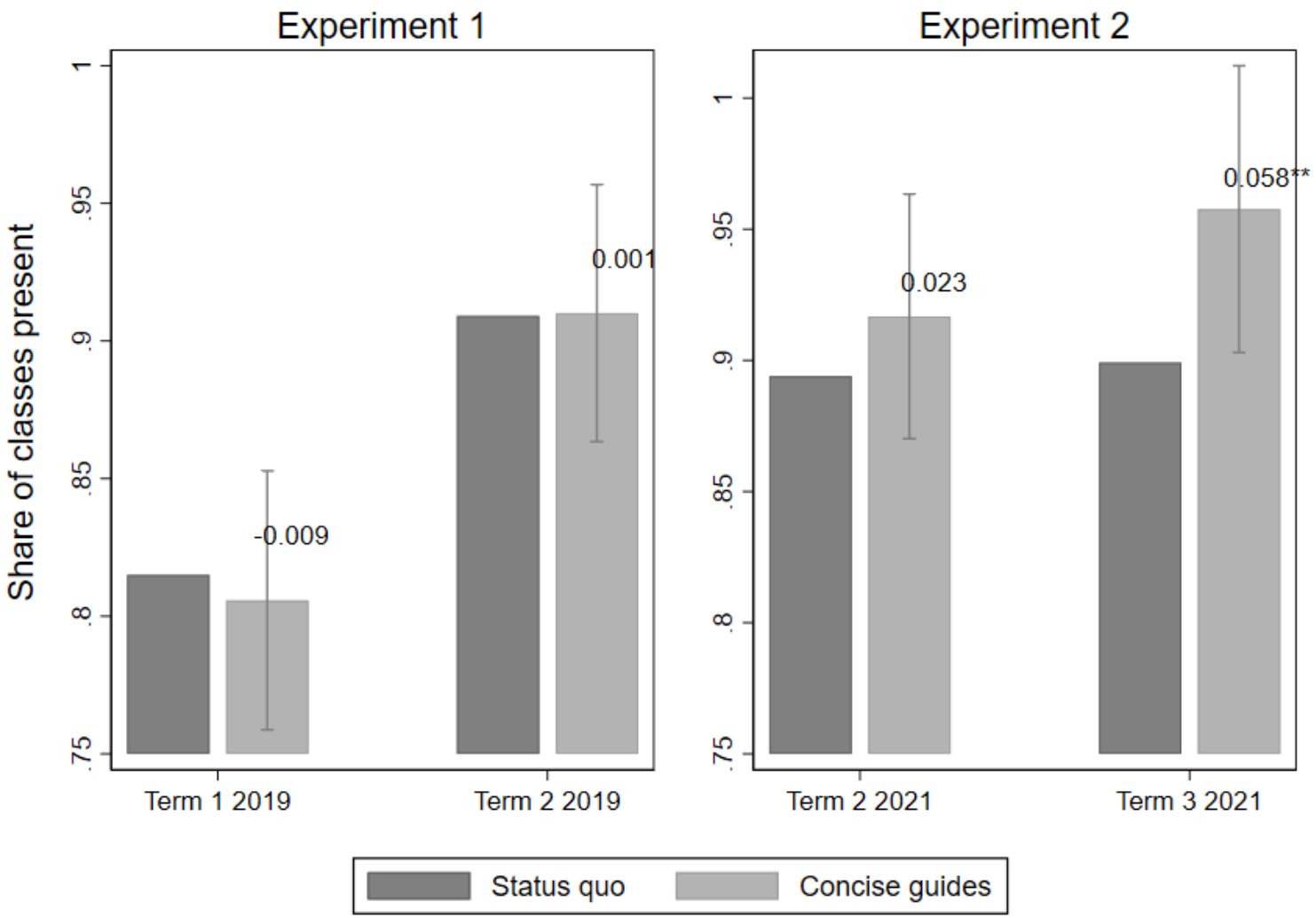


Figure 6: Notes: Figure shows the impact of less-detailed guides on teacher attendance. Effect size is reported above the less-detailed guide bar. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table 1: Textual analaysis
Experiment 2

	More detailed guides (1)	Less detailed guides (2)	Difference (3)
Panel A: Total Sentences			
Number of words	972.66	469.92	502.74
			[0.00]
Number of sentences	117.81	62.68	55.13
			[0.00]
Average words per sentence	8.34	7.46	0.88
			[0.00]
Readability score	80.72	80.42	0.30
			[0.46]
Panel B: Read-aloud sentences			
Number of words	387.38	200.79	186.59
			[0.00]
Number of sentences	40.70	25.95	14.75
			[0.00]
Average words per sentence	9.19	7.51	1.68
			[0.00]
Readability score	80.46	80.87	-0.41
			[0.38]
Panel C: Instructions to teachers			
Number of words	585.28	269.13	316.15
			[0.00]
Number of sentences	77.12	36.73	40.39
			[0.00]
Average words per sentence	7.79	7.14	0.65
			[0.00]
Readability score	80.50	80.83	-0.33
			[0.54]
Panel D: Boxed Answer Keys and Board Content			
% of guides with boxed answer keys and board content(including images and texted boxed content)	0.58	0.84	-0.26
			[0.00]

Notes: This table presents a comparative text analysis among status quo and concise guides in Experiment 2. All statistics provided are guide-level comparisons. Panel A reports metrics for the entire text, including the number of sentences, total word count, average words per sentence, and readability scores. Panel B focuses on these metrics exclusively for bolded sentences(intended to be read out loud to the class), while Panel C provides a similar analysis for unbolded sentences (for the teacher's reference only). Panel D compares the proportion of guides featuring boxed content, which encompasses answer keys and board content. *P-values* from a test that the forms are equivalent on each dimension are reported in brackets. Readability is assessed using the Flesch-Kincaid readability test scores. Higher scores indicate greater ease of reading.

Table 2: Characteristics of schools, balance, and follow-up
Experiment 1

	More detailed guides		Less detailed guides				Observations	
	Mean (1)	SD (2)	Mean (3)	SD (4)	Difference (5)	p-value (6)	Academies (7)	Pupils (8)
Panel A: Follow-up rates for test score outcomes								
Midline 1	0.72	0.45	0.76	0.43	0.03	0.549	184	2,917
Midline 2	0.57	0.50	0.63	0.48	0.06	0.147	184	2,917
Endline	0.59	0.49	0.59	0.49	0.00	0.988	184	2,917
Panel B: Baseline characteristics of sample								
Baseline math	0.05	0.93	0.18	0.88	0.12	0.152	172	1,727
Baseline reading	0.09	0.92	0.18	0.89	0.07	0.375	162	1,599
Baseline Kiswhaili	0.11	0.94	0.04	0.98	-0.07	0.717	102	980
Female	0.48	0.50	0.48	0.50	0.00	0.931	172	1,726
Years enrolled at Bridge	2.77	1.51	2.67	1.50	-0.12	0.263	172	1,727
Age	8.40	0.96	8.41	0.86	0.02	0.786	172	1,724
Enrollment (pupil-teacher ratio)	20.05	8.90	20.13	8.36	0.60	0.713	172	1,727
Urban/peri-urban location	0.48	0.50	0.56	0.50	0.08	0.469	114	990

Notes: Sample size varies across covariates due to data availability. Estimates of the difference between treatment and comparison also control linearly for the probability of treatment in the randomization stratum. The p-value accounts for errors clustering at the academy level.

Table 3: Follow-up, descriptive statistics, and covariate balance
Experiment 2

	More detailed guides		Less detailed guides				Observations	
	Mean (1)	SD (2)	Mean (3)	SD (4)	Difference (5)	p-value (6)	Academies (7)	Pupils (8)
Panel A: Follow-up rates for test score outcomes								
Midline 1	0.76	0.43	0.84	0.37	0.08	0.167	112	1,962
Midline 2	0.77	0.42	0.81	0.39	0.04	0.104	112	1,962
Endline	0.81	0.39	0.78	0.42	-0.04	0.430	112	1,962
Panel B: Baseline characteristics of sample								
Baseline language score	0.05	1.03	0.15	1.02	0.10	0.389	103	1,304
Baseline composition score	0.09	0.93	0.13	1.05	0.04	0.797	102	1,306
Female	0.50	0.50	0.52	0.50	0.02	0.356	107	1,551
Years enrolled at Bridge	4.36	2.35	4.41	2.34	0.05	0.777	107	1,551
Age	12.44	0.95	12.47	1.08	0.03	0.653	107	1,548
Enrollment (pupil-teacher ratio)	24.58	9.01	23.77	7.67	-0.81	0.630	106	1,547
Urban/peri-urban location	0.76	0.42	0.58	0.49	-0.19	0.015	107	1,560

Notes: Sample size varies across covariates due to data availability. Panel B restricts to the sample of students with endline test scores. The p-value accounts for errors clustered at the academy level.

Table 4: Effects of less detailed teacher guides on math test scores
Experiment 1

	(1)	(2)	(3)	(4)	(5)
Less detail	0.168**	0.086	0.173***	0.189***	0.185***
	(0.073)	(0.072)	(0.074)	(0.072)	(0.071)
Class size		-0.003			
		(0.006)			
Less detail × Class size		0.016**			
		(0.008)			
Baseline score	0.472***	0.475***	0.498***		
	(0.030)	(0.029)	(0.046)		
Less detail × Baseline score			-0.057		
			(0.055)		
Residualized baseline score				0.530***	0.531***
				(0.043)	(0.028)
Less detail × Residualized score				0.002	
				(0.056)	
Classroom mean baseline score				0.313***	0.409***
				(0.086)	(0.130)
Less detail × Classroom mean score					-0.206
					(0.153)
Observations	3,527	3,527	3,527	3,527	3,527
Number academies	174	174	174	174	174
R^2	0.23	0.23	0.23	0.24	0.24

Notes: This table shows the effects of reducing the teacher guide length on post-treatment math test scores. All specifications control for baseline math scores. The PTR variable is centered on the median PTR across students (15) so that the main effect represents the effect for the median student in terms of PTR. A standard deviation of the class size distribution across students is 4.7. A standard deviation of the residualized baseline test score is 0.860. A standard deviation of the mean baseline test score is 0.053. All specifications control linearly for the probability of treatment in the school's randomization stratum. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table 5: Effects of less detailed teacher guides on math test scores
Experiment 2

	(1)	(2)	(3)	(4)	(5)
Less detail	0.170*	0.088	0.159*	0.175**	0.156**
	(0.089)	(0.087)	(0.085)	(0.084)	(0.074)
Class size		-0.010			
		(0.010)			
Less detail × Class size		0.030***			
		(0.012)			
Baseline score	0.605***	0.602***	0.561***		
	(0.030)	(0.029)	(0.047)		
Less detail × Baseline score			0.106		
			(0.068)		
Residualized baseline score				0.591***	0.614***
				(0.048)	(0.032)
Less detail × Residualized baseline score				0.059	
				(0.064)	
Classroom mean baseline score				0.507***	0.298*
				(0.168)	(0.154)
Less detail × Classroom mean baseline score					0.423
					(0.292)
Observations	1,304	1,304	1,304	1,304	1,304
Number academies	103	103	103	103	103
R^2	0.503	0.510	0.506	0.505	0.510

Notes: This table shows the effects of reducing the teacher guide length on math scores. The PTR variable is centered on the median PTR across students (17.5) so that the main effect represents the effect for the median student in terms of PTR. A standard deviation of the class size distribution across classrooms is 7.2. A standard deviation of the residualized baseline test score is 0.858. A standard deviation of the mean baseline test score across classrooms is 0.539. All specifications randomization strata fixed effects. Standard errors are clustered at the strata level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table 6: Dispersion effects
Experiment 1

	Overall		Within classrooms		Between classrooms	
	More detailed guides (1)	Less detailed guides (2)	More detailed guides (3)	Less detailed guides (4)	More detailed guides (5)	Less detailed guides (6)
Panel A: Below median class size						
Standard deviation	0.84	0.80	0.66	0.68	0.52	0.42
Difference	-0.04		0.02		-0.10	
F-test p-value	0.45		0.70		0.21	
Levene's test p-value	0.42		0.49		0.21	
Observations	451		451		79	
Panel B: Above median class size						
Standard deviation	1.06	0.91	0.87	0.80	0.64	0.44
Difference	-0.15		-0.07		-0.20	
F-test p-value	0.00		0.04		0.02	
Levene's test p-value	0.00		0.04		0.04	
Observations	1,276		1,276		93	

Notes: This table compares the dispersion of test scores between within cells formed by experimental condition and classroom PTR. Panel A shows the dispersion in test scores for classrooms below the median level of class size. Panel B shows the same for classrooms above or equal to the median.

Table 7: Dispersion effects
Experiment 2

	Overall		Within classrooms		Between classrooms	
	More detailed guides (1)	Less detailed guides (2)	More detailed guides (3)	Less detailed guides (4)	More detailed guides (5)	Less detailed guides (6)
Panel A: Below median class size						
Standard deviation	0.90	0.89	0.78	0.76	0.46	0.55
Difference		-0.01		-0.02		0.09
F-test p-value		0.91		0.76		0.40
Levene's test p-value		0.77		0.67		0.49
Observations		448		448		48
Panel B: Above median class size						
Standard deviation	0.93	1.00	0.86	0.81	0.37	0.56
Difference		0.07		-0.05		0.19
F-test p-value		0.08		0.13		0.03
Levene's test p-value		0.08		0.38		0.19
Observations		1,112		1,112		59

Notes: This table compares the dispersion of test scores between within cells formed by experimental condition and classroom PTR. Panel A shows the dispersion in test scores for classrooms below the median level of class size. Panel B shows the same for classrooms above or equal to the median.

Table 8: Effects on teacher behaviors

	Teacher attendance		Lesson completion		Completion conditional on attendance		Total instruction time (hours)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Experiment 1								
Less detail	0.001	0.001	-0.039	-0.039	-0.006	-0.006		
	(0.024)	(0.024)	(0.029)	(0.028)	(0.057)	(0.057)		
Class size		-0.000		-0.002		-0.003		
		(0.001)		(0.003)		(0.003)		
Less detail × class size		0.000		-0.004		-0.007		
		(0.002)		(0.004)		(0.005)		
Observations	178		178		176			
Comparison mean of dep var	0.909		0.667		0.740			
Panel B: Experiment 2								
Less detail	0.058**	0.062**	0.084**	0.086**	0.061	0.062	3.232*	3.297*
	(0.028)	(0.029)	(0.040)	(0.041)	(0.041)	(0.042)	(1.790)	(1.864)
Class size		0.003		-0.000		-0.001		0.020
		(0.004)		(0.007)		(0.007)		(0.292)
Less detail × class size		0.001		0.003		0.002		0.048
		(0.007)		(0.011)		(0.011)		(0.482)
Observations	112		111		111		111	
Comparison mean of dep var	0.899		0.667		0.715		30.212	

Notes: Table reports on effects of teacher behavior in the final period of each experiment. Teacher attendance is recorded by academy managers. Lesson completion is recorded telemetrically using the lesson guide. Lesson completion indicates that the teacher reached the end of a lesson. Total instruction time is the amount of time that the teacher was recorded as having a lesson open on their tablet. Completion conditional on teacher attendance rescales the completion rate by the teacher attendance rate. Robust standard errors reported in parentheses. Specifications in Panel A control for the classroom's probability of assignment to the experimental condition. Panel B specifications control for randomization strata fixed effects. ***, **, and * indicate significance at 1%, 5%, and 10%.

Appendices

Appendix A Data Documentation

Baseline test scores. For Experiment 1, NewGlobe supplied three baseline test scores in math, English reading, and Kiswahili observed prior to randomization: the Term 1 midterm and endterm and the Term 2 midterm. Students may be missing particular baseline test scores.

We construct a single score representing the pupil's baseline level in each subject by standardizing and averaging all non-missing test scores.

In Experiment 2, NewGlobe shared baseline test scores in English language and composition from Term 1 midterm from 2020, the last time the students would have been tested before the start of the intervention. Throughout we use the standardized baseline language score.

Class size. We measure class size as the number of unique student identifiers associated with each classroom who have test scores in the baseline file. NewGlobe also supplied records of pupil-teacher ratios in each school. In the first experiment, we have NewGlobe's record of PTR prior to baseline. In the second experiment, the record is from after the start of the intervention, raising concerns that these may be endogenous. To have a symmetric way of calculating class size across the experiments, we have the number of unique students found in the baseline test score data. Results are very similar if we use the PTR records instead of the number of unique student identifiers in the baseline test score data.

Teacher behavior. NewGlobe provided data on teacher behaviors, including attendance and lesson completion. Teacher attendance data reflect the fraction of class sessions at which the classroom's regular teacher attended. Bridge employs substitute teachers when teachers are absent, so effects on teacher attendance may not reflect increases in the amount of instructional time students receive.

Lesson completion data consists of the number of scheduled lessons and the number of lessons "completed" by the teacher in each academic term.[20] In the first experiment, these data are

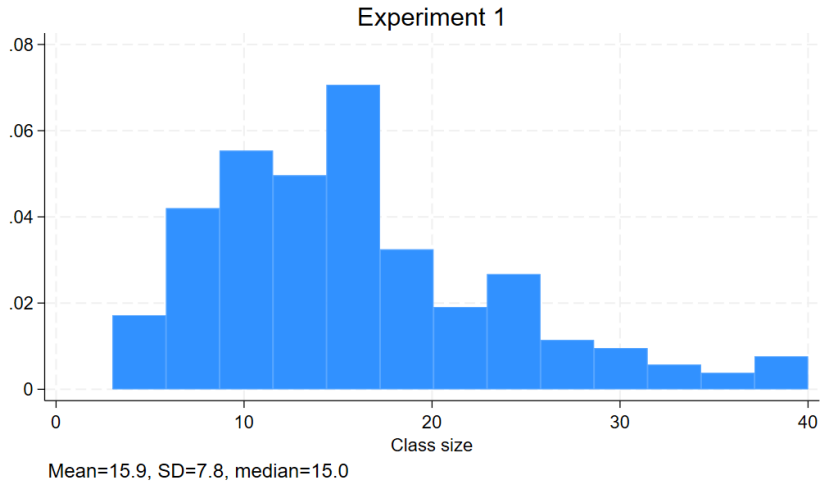
[20]The data also include information on the number of lessons opened. However, we find that this measure is highly predictive of completion, and therefore do not report these results for brevity.

disaggregated by the particular instructional component: activities and revision. We use these data to construct a lesson completion index that averages the percentage of total scheduled lessons scheduled for both components. In the second experiment, NewGlobe shared data at the individual lesson level, including the amount of time spent delivering the lessons. We use these data to calculate the fraction of lessons completed. We are also able to calculate the number of instructional hours that students receive based on these data.

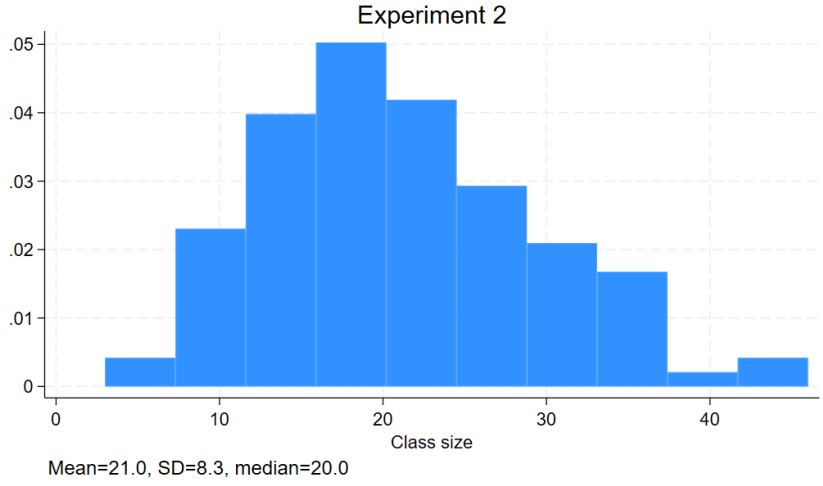
It should be noted that lesson completion indicates the fraction of lessons *completed by the classroom's regular teacher*. A teacher who is present more often will complete a larger share of lessons according to this measure. This complicates the interpretation of effects on lesson completion where there is a corresponding change in teacher attendance, as it is possible that the intervention only changed the fraction of lessons completed by the regular classroom teacher.

Appendix B Figures

Figure A1



(a)



(b)

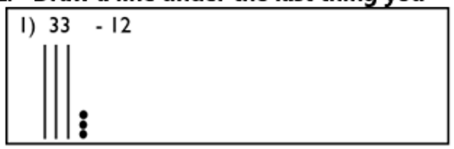
Notes: Histograms of class size. Each figure shows the distribution of class size across pupils.

Figure A2

More detailed guide (comparison)

My Turn: 5 Minutes

1. Write on the board:
2. **Draw a line under the last thing you**



wrote and write today's date.

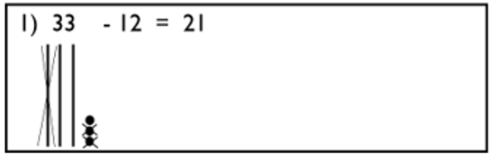
3. Scan.
4. **Problem 1 says 33 - 12.**
5. **What does problem 1 say, class?** [Signal] *33 - 12.*
6. Point to the 3 lines and 3 dots.
7. **I drew 3 tens and 3 ones.**
8. **This shows 33.**
9. Point to – 12.
10. **Now let's take away 12.**
11. **12 has 1 ten and 2 ones.**
12. **That means I am taking away 1 ten and 2 ones from 33.**
13. Add to the board:



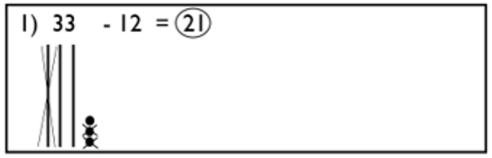
14. **First I cross out 1 ten.**
15. Add to the board:



16. **Now I cross out 2 ones.**
17. **I had 33 and I took 12 away.**
18. **Let's count what is left.**
19. Point to the line.
20. **I have 2 tens.**
21. Point to the dot.
22. **I have 1 one.**
23. **2 tens and 1 one makes 21.**
24. **This means I have 21 left.**
25. Add to the board:



26. **Therefore, 33 - 12 is 21.**
27. Add to the board:



28. **This is my final answer.**

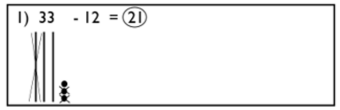
Less detailed guide (treatment)

My Turn: 5 Minutes

1. Write on the board:



2. Demonstrate how to solve the problem on the board. See below for completed problem.



Notes: Comparison of teacher guides Experiment 1.

Figure A3

Version A Teacher Guide Lesson 3

7. **Class, a NOUN is a person, place, thing or idea. There are three types of nouns.**
8. **A SINGULAR NOUN names ONE person, place, thing or idea. What does a singular noun name?** [Signal]*One person, place, thing, or idea.*
9. **"Girl," "park," "pencil," and "love" are examples of singular nouns.**
10. **A PLURAL NOUN names MORE THAN ONE person, place, thing, or idea. What does a plural noun name?** [Signal] *More than one person, place, thing, or idea.*
11. **"Girls," "parks," "pencils," and "books" are examples of plural nouns. They all end with the letter s.**
12. **A PROPER NOUN names a SPECIFIC person, place, thing, or idea. What does a proper noun name?** [Signal] *A specific person, place, thing, or idea.*
13. **"Nelson Mandela," "Kenya," "October," and "Ramadan" are all examples of proper nouns. Proper nouns ALWAYS begin with a capital letter.**

Version B Teacher Guide Lesson 3

7. **I will read aloud the introduction while you follow along in your textbook.**
8. Read aloud the lesson's introduction.

Textbook Lesson 3

Lesson 3 Grammar: Nouns and Proper Nouns

A **singular** noun names one person, place, thing, or idea.

A **plural** noun names more than one person, place, or thing by adding –s.

Examples

- There are 5 boy**s** in the class.
- He owns many farm**s**.
- She lost all of her pencil**s**.

People	Places	Things	Ideas
girl	park	pencil	love
teacher	school	book	hate
boy	forest	tree	happiness
mother	farm	ball	freedom
uncle	restaurant	cloud	courage
doctor	market	cup	sadness

Notes: Comparison of teacher guides from Experiment 2. Version A corresponds to the more detailed version, and Version B less detailed. This example illustrates how the treatment guides instruct the teacher to read from the textbook, instead of reading the instructions directly from the guide. Note that in both cases, students would have been instructed to open their textbooks to the relevant page.

Appendix C Tables

Table A1: Simplicity rules

	Rule	Definition
1	Keep it linear	-Don't mention step B before you do step A. -Don't set the teacher up to scroll backwards.
2	Use lean language	-Don't say it in four words if you can say it just as clearly in two. -Avoid non-essential words
3	Use consistent language	-Avoid slight variations in the way that you script repeated instructions.
4	No optionality	-Do not script conditional scenarios -Do not provide options
5	No multi-step instructions	-When pupils are working don't direct the teacher to do four things, direct the teacher to do the one most important thing
6	Use teacher training language properly	-Use the 'approved' language exactly as is, don't make slight variations -Use it sparingly and only in the highest leverage places

Table A2: Attrition on individual test scores
Experiment 1

	More detailed guides		Less detailed guides				Observations		
	Mean	SD	Mean	SD	Difference	p-value	Academies	Pupils	Date
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
2018 Term 3									
Posttest: Math activities	0.75	0.44	0.81	0.39	0.06	0.233	184	2,917	Oct 18, 2018
Posttest: Math revision	0.75	0.43	0.77	0.42	0.02	0.647	184	2,917	Oct 18, 2018
Endterm	0.67	0.47	0.70	0.46	0.04	0.314	184	2,917	Oct 23, 2018
2019 Term 1									
Pretest: Math activities	0.62	0.49	0.52	0.50	-0.10	0.048	184	2,917	Jan 7, 2019
Pretest: Math revision	0.64	0.48	0.53	0.50	-0.11	0.018	184	2,917	Jan 7, 2019
Midterm	0.69	0.46	0.74	0.44	0.04	0.499	184	2,917	Feb 14, 2019
Posttest: Math activities	0.57	0.49	0.64	0.48	0.07	0.122	184	2,917	Mar 25, 2019
Posttest: Math revision	0.62	0.49	0.64	0.48	0.03	0.548	184	2,917	Mar 25, 2019
Endterm	0.72	0.45	0.70	0.46	-0.01	0.810	184	2,917	
2019 Term 2									
Pretest: Math activities	0.58	0.49	0.64	0.48	0.06	0.124	184	2,917	May 6, 2019
Pretest: Math revision	0.59	0.49	0.61	0.49	0.02	0.647	184	2,917	May 6, 2019
Midterm	0.62	0.48	0.62	0.49	-0.01	0.827	184	2,917	Jun 12, 2019
Posttest: Math activities	0.60	0.49	0.61	0.49	0.01	0.765	184	2,917	Jul 23, 2019
Posttest: Math revision	0.60	0.49	0.60	0.49	0.00	0.992	184	2,917	Jul 23, 2019
Endterm	0.65	0.48	0.65	0.48	-0.01	0.836	184	2,917	Jul 30, 2019

Notes: The p-value accounts for errors clustering at the academy level.

Table A3: Attrition on individual test scores
Experiment 2

	More detailed guides		Less detailed guides				Observations		
	Mean	SD	Mean	SD	Difference	p-value	Academies	Pupils	Date
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Term 2									
Language endterm	0.76	0.43	0.84	0.37	0.08	0.251	112	1,962	March 15, 2021
Composition endterm	0.74	0.44	0.79	0.40	0.05	0.497	112	1,962	March 15, 2021
Fluency endterm	0.73	0.44	0.81	0.39	0.08	0.295	112	1,962	March 17, 2021
Term 3									
Language midterm	0.77	0.42	0.81	0.39	0.04	0.247	112	1,962	June 10, 2021
Composition midterm	0.77	0.42	0.81	0.39	0.04	0.244	112	1,962	June 10, 2021
Fluency midterm	0.77	0.42	0.81	0.39	0.04	0.298	112	1,962	June 10, 2021
Language endterm	0.81	0.39	0.78	0.42	-0.04	0.458	112	1,962	July 13, 2021
Composition endterm	0.81	0.39	0.79	0.40	-0.02	0.667	112	1,962	July 13, 2021
Fluency endterm	0.75	0.43	0.68	0.47	-0.07	0.328	112	1,962	July 15, 2021

Notes: The p-value accounts for errors clustering at the academy level.

Table A4: Effects on all test scores
Experiment 1

	Math revision		Math activities			
	Pretest	Posttest	Pretest	Posttest	Midterm	Endterm
	(1)	(2)	(3)	(4)	(5)	(6)
Term 3 2018						
Less detail effect	-0.07	0.03	0.22	-0.00	0.03	0.03
	(0.13)	(0.08)	(0.16)	(0.10)	(0.07)	(0.07)
	1,411	2,217	1,589	2,264	1,644	1,999
Term 1 2019						
Less detail effect	0.11	0.12	0.12	0.08	-0.01	0.10
	(0.11)	(0.09)	(0.13)	(0.12)	(0.09)	(0.09)
	1,714	1,839	1,666	1,772	2,083	2,078
Term 2 2019						
Less detail effect	0.21**	0.14*	-0.00	0.19**	0.07	0.05
	(0.09)	(0.08)	(0.11)	(0.08)	(0.06)	(0.07)
	1,756	1,758	1,779	1,769	1,809	1,901

Notes: Table shows the effect of less detailed guides on each post-intervention test score. All specifications control linearly for the probability of assignment to the concise guide condition within the school's randomization strata and the student's baseline math performance index. Standard errors reported in parentheses are clustered at the school level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A5: Effects on all test scores
Experiment 2

	Language (1)	Composition (2)	Fluency assessment (3)
Term 2 endterm			
Less detail effect	0.09	-0.09	0.06
	(0.11)	(0.16)	(0.18)
	1,315	1,271	1,278
Term 3 midterm			
Less detail effect	0.15	-0.01	-0.03
	(0.10)	(0.13)	(0.17)
	1,324	1,332	1,333
Term 3 endterm			
Less detail effect	0.17*	-0.00	-0.07
	(0.09)	(0.14)	(0.19)
	1,304	1,320	1,174

Notes: Table shows the effect of less detailed guides on all post-intervention test score. The primary test score outcome was the language assessment. All specifications control for baseline language scores and randomization strata fixed effects. Standard errors reported in parentheses are clustered at the randomization strata level. ***, **, and * indicate significance at 1%, 5%, and 10%.

Table A6: Relationship between teacher lesson completion data and teacher attendance

	Lesson completion rate		Total instructional time (hours)		Mean time per lesson (minutes)	
	(1)	(2)	(3)	(4)	(5)	(6)
[1em] Teacher attendance rate	1.102***	1.073***	51.26***	49.84***	-2.979	-2.627
	(0.123)	(0.132)	(4.070)	(4.488)	(7.717)	(7.639)
Less detailed guides		0.0366		1.843		-0.273
		(0.0311)		(1.274)		(0.719)
Observations	111	111	111	111	108	108

Notes: Table reports the OLS relationship between teacher attendance and measures of teacher lesson completion from the final term of Experiment 1. Lesson completion is observed only if the teacher is present, even if a substitute teacher completes lessons.

Table A7: Effects of less detailed teacher guides on math test scores
By gender

	Experiment 1 (1)	Experiment 2 (2)
Less detail	0.123	0.199*
	(0.082)	(0.102)
Female	0.049	-0.036
	(0.051)	(0.051)
Less detail guides × Female	0.095	-0.057
	(0.072)	(0.066)
Observations	3,525	1,298
Number academies	174	103
R^2	0.23	0.51

Notes: This table shows the effects of reducing the teacher guide length on post-treatment math test scores. ***, **, and * indicate significance at 1%, 5%, and 10%.