ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-586

Assessors influence results: Evidence on enumerator effects and educational impact evaluations

Daniel Rodriguez-Segura
University of Virginia

Beth E. Schueler
University of Virginia

A significant share of education and development research uses data collected by workers called "enumerators." It is well-documented that "enumerator effects"—or inconsistent practices between the individual people who administer measurement tools— can be a key source of error in survey data collection. However, it is less understood whether this is a problem for academic assessments or performance tasks. We leverage a remote phone-based mathematics assessment of primary school students and survey of their parents in Kenya. Enumerators were randomized to students to study the presence of enumerator effects. We find that both the academic assessment and survey was prone to enumerator effects and use simulation to show that these effects were large enough to lead to spurious results at a troubling rate in the context of impact evaluation. We therefore recommend assessment administrators randomize enumerators at the student level and focus on training enumerators to minimize bias.

VERSION: June 2022

Suggested citation: Rodriguez-Segura, Daniel, and Beth Schueler. (2022). Assessors influence results: Evidence on enumerator effects and educational impact evaluations. (EdWorkingPaper: 22-586). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/hfgf-3404

DRAFT AS OF JUNE 2022

Assessors influence results: Evidence on enumerator effects and educational impact evaluations

Daniel Rodriguez-Segura
dan.rodriguez@virginia.edu
University of Virginia, School of Education and Human Development

Beth E. Schueler
beth_schueler@virginia.edu
University of Virginia, School of Education and Human Development

Abstract: A significant share of education and development research uses data collected by workers called "enumerators." It is well-documented that "enumerator effects"—or inconsistent practices between the individual people who administer measurement tools—can be a key source of error in survey data collection. However, it is less understood whether this is a problem for academic assessments or performance tasks. We leverage a remote phone-based mathematics assessment of primary school students and survey of their parents in Kenya. Enumerators were randomized to students to study the presence of enumerator effects. We find that both the academic assessment and survey was prone to enumerator effects and use simulation to show that these effects were large enough to lead to spurious results at a troubling rate in the context of impact evaluation. We therefore recommend assessment administrators randomize enumerators at the student level and focus on training enumerators to minimize bias.

Keywords: enumerator effects, impact evaluations, remote assessments, education in developing countries

Author's note: Daniel Rodriguez-Segura is the corresponding author. We are grateful for feedback received from Steven Glazerman, Betheny Gross, Jim Soland, Daphna Bassok, Vivian Wong and Isaac Mbiti, and our colleagues in the Center on Education Policy and Workforce Competitiveness (EdPolicyWorks). We also thank Tim Sullivan, Veronica Kimani, Sean Geraghty, and others at NewGlobe for partnering with us on this project as well as Lee Crawfurd for generously sharing the assessments used in Crawfurd et al. (2021), and Shannon Kontalonis for excellent administrative support. This work has been supported by Innovations for Poverty Action (IPA) through the Research Methods Initiative (grant code NWU0004-X1), and the Center for Reinventing Public Education (CRPE). The authors received IRB approval the University of Virginia, protocol number 3751. This trial was pre-registered at the AEA RCT Registry (number AEARCTR-0006913) after the data collection was completed but before analysis and transfer of data. Declarations of interest: Daniel Rodriguez-Segura has accepted a job at NewGlobe, the research partner for this project. Daniel's new role started after the analysis of the data and the writing of this paper.

I. Introduction

A significant share of research in low- and middle-income countries (LMICs) relies on data collected directly and on a one-on-one basis by workers called "enumerators"[1] (Lupu and Michelitch, 2018). This is particularly true in education research where surveys of teachers, parents, school leaders, and one-on-one assessments of student academic achievement are often conducted by teams of assessors. When collecting this type of data, individual enumerators can influence the recruitment of subjects, exercise significant discretion in how they interpret the information received, and shape how responses end up being coded in the data. In doing so, assessors can affect the quality of the data received by introducing measurement error through "enumerator effects", which happen when assessors record differential response rates or scores for similar populations of respondents (Olson et al., 2020). These inconsistent practices across enumerators or systematic variation in how respondents react to different enumerators can lead to erratic data, which could in turn yield spurious research results and unhelpful policy recommendations. Although the presence of enumerator effects has been well-documented in survey data (West and Blom, 2017; Di Maio and Fiala, 2018), researchers have not yet studied as much the extent to which they can affect educational assessments and the results of impact evaluations using these data.

In this paper, we study the presence of enumerator effects in a learning assessment of primary school children's early numeracy skills and an accompanying survey of their parents for over 2,500 students across 105 schools in Kenya, delivered by 20 individual assessors. We leverage the fact that this assessment was centrally administered over the telephone, removing many logistical or geographic barriers to creating a fully interpenetrated design. In other words, administering this assessment and survey remotely allowed us to randomly assign enumerators to students at the individual-level. Importantly, this phone-based assessment (PBA) was the first set of outcomes collected after a randomized impact evaluation of a remote mathematics instructional intervention implemented in Kenya while in-person schooling was on hold due to the COVID-19 pandemic (Schueler and Rodriguez-Segura, 2021). This enables us to also estimate how much any enumerator effects could have biased the estimation of treatment effects in this impact evaluation had enumerators been assigned in a more typical manner. Finally, this PBA also included more traditional survey questions directed at parents, which lets us compare enumerator effects on educational assessments to those on survey questions, in a realm closer to the type of measures for which other researchers have previously studied enumerator effects and which are commonly used by social scientists within and beyond the field of education.

[1] Enumerators are sometimes referred to as "assessors" or "interviewers". We will use these terms interchangeably.

Motivation and Contribution

One-on-one surveys and assessments that are administered by teams of assessors are prevalent in education research. Questionnaires that are widely-used in research and policy planning like national censuses or Demographic and Health Surveys (DHS), and internationally-validated exams like the Early Grade Reading Assessment (EGRA) — adapted for over 65 countries (Dubeck and Grove, 2015) — or the nationally representative Annual Status of Education Report (ASER) in India follow this approach. Many important studies and reports have leaned on this type of data for either the framing of their research questions, or as direct outcome measures (e.g., Mbiti et al., 2019; Evans and Mendez Acosta, 2021; World Bank, 2018, Varly, 2020). However, this type of data collected by teams of individual assessors on a one-on-one basis, is susceptible to enumerator effects, or the non-zero correlations among the responses collected by an individual assessor or interviewer (West and Blom, 2017.)

Enumerator effects could play a significant role in shaping assessment results and indeed the state of the literature as a whole through several potential channels. First, individual assessors could exercise more or less leniency in what is considered a "correct" answer, or they could vary how long children are actually given to answer questions. The level of leniency could be impacted, in part, by the assessor's familiarity with the content or the individual student being assessed. For instance, if an enumerator that is unfamiliar with fractions is asking a question on this topic and the answer sheet says that the correct answer for this question is "1/4", they might mark an answer of "0.25" wrong. Some enumerators may be more forthcoming than others when students ask for clarification about a given question in ways that systematically advantage or disadvantage the students they assess. Additionally, the same respondent may provide different answers depending on their perceptions of a given enumerator. For example, a student may be less likely to try hard on an academic assessment if the assessor is a stranger than a known teacher. Similarly, on survey-based measures, parents may be less likely to divulge personal information if they perceive the interviewer to be indiscrete. Even subtle differences in tone when reading a question could influence the response.

One potential negative implication of enumerator effects is that they could influence the accuracy and precision of estimates comparing between groups, such as a treatment effect estimate in the context of impact evaluation. Mechanically, higher intra-enumerator correlations increase the error introduced into these estimates through higher variance in the responses (Olson et al. 2020). Therefore, the presence of enumerator effects can result in unwanted inflation of the estimate's variance, ultimately reducing the statistical power and precision of the results. Increased variance could influence the accuracy of the point estimates in a study by allowing for a wider range of probable averages for the treatment and control groups individually due to higher measurement error within each. This would

in turn increase the probability of observing a difference (or lack of thereof) between the treatment and comparison groups that was simply due to the measurement error introduced by the enumerator effects, and not as a result of a "true" treatment effect. In fact, beyond the literature on enumerator effects, researchers have documented cases in which measurement error in survey data was introduced systematically differently for treatment and comparison groups (Baird and Özler, 2012; Blattman et al., 2016), potentially leading to biased estimated treatment effects if the researchers relied solely on survey data. Therefore, given the wider variance introduced by enumerator effects, and the previous documentation of how systematic differences in measurement errors across treatment and control groups might lead to spurious results in impact evaluations, it is reasonable to suspect that enumerator effects in survey and assessment data could lead to biased treatment estimates. Yet, the extent to which enumerator effects bias impact estimates is currently a gap in our empirical knowledge.

Another related concern introduced by enumerator effects is that the assignments of enumerators to clusters of respondents (e.g., classrooms, schools, districts) could lead to biased point estimates because of "unlucky bunching" in one group over the other (e.g., in the treatment versus control group in the context of an impact evaluation). For example, imagine an experiment where there are only two schools (one in each experimental group), two assessors that will each be assigned to one school, and a "true" treatment effect that is null. If there are no systematic differences in how these enumerators record answers, then the estimate of the difference between the treatment and control schools (i.e., the estimated treatment effect) will be accurately null. However, if one of the assessors records systematically higher scores than the other, then the treatment effect will be either positive or negative, even if there is no true treatment effect. In this case, enumerator effects combined with the assignment of assessors at the school level will lead to biased estimates. Yet, even in the presence of enumerator effects, the less clustered the assignment of the assessors is, the less bias that one would expect enumerator effects to introduce into the treatment effect. For instance, in the previous example, if both assessors were randomly assigned to half of the students in both schools, the treatment effects would cancel out, and one would end up with the correct treatment estimate. In practice, due to logistical constraints, assessors are often assigned to clusters in such a way that could lead researchers to generate biased estimates in cases where enumerator effects are operating.

In the case of surveys, the presence of enumerator effects has been long known to social scientists, including in the context of developing countries (West and Blom, 2017), along three different lines of work. First, enumerator effects have been documented as a result of different observable characteristics of enumerators, subjects, or the interaction of both. For example, Adida et al. (2016) find evidence of enumerator effects in large-scale surveys across African 14 countries, especially when the ethnic group of the respondent and

interviewer have a history of political competition. Similarly, Di Maio and Fiala (2018) find that in Uganda, although most observable characteristics of assessors yield minimal enumerator effects, when enumerators are asking highly sensitive political preference questions, differences between enumerators account for over 30 percent of the variation in responses. Benstead (2014) and Blaydes and Gillum (2013) find that the perceived religiosity of the interviewer affects response patterns in surveys administered in Morocco and Egypt respectively, as respondents provided more "socially desirable" answers depending on the appearance of the interviewer. Secondly, enumerator effects have also been documented due to enumerators interpreting the content of questionnaires differently. For instance, Randall et al. (2013) show that the word "household" is difficult to translate into some languages, leading enumerators to venture into their own conceptual interpretation of the questions when presenting them to respondents, and as such, increases the potential for wider variance in the data. Finally, enumerator effects have also been documented appearing at different rates depending on the content of the survey. For example, Himelein (2016) finds interviewer effects in a survey in Timor Leste across subjective and objective questions, but with effects of larger size for subjective questions.

Although enumerator effects have been well studied in the context of survey research, there has been much less formal documentation of enumerator effects in educational assessments. This is true despite the fact that the characteristics of one-on-one educational assessments also make them susceptible to enumerator effects in that, they are often administered by teams of assessors assigned to assess clusters of students, and require discretion from assessors to mark questions, to properly allocate how much time allow for each question, to provide clarification to students on confusing questions, and to know when to stop an assessment. Part of the difficulty of studying enumerator effects in education, and in development research more broadly, is logistical (Di Maio and Fiala, 2018; West and Blom, 2017). Specifically, to clearly isolate the extent of "enumerator effects" during a round of data collection, researchers would ideally create "fully interpenetrated designs" where the assignment of assessors to respondents is randomized at the individual-level[2] (West and Blom, 2017). In other words, fully interpenetrated designs randomly assign individual subjects to enumerators so that, in expectation, any major differences in response patterns obtained by individual enumerators would be due to the enumerators themselves and not due to differences in the respondent pool assigned to each enumerator.

In spite of the methodological desirability of fully-interpenetrated designs, the physical logistics of randomizing assessors to students at the individual-level can be

[2] In the absence of fully-interpenetrated designs, researchers in the past have had to instead use complicated hierarchical statistical models to isolate enumerator effects, ranging from basic random effects models, to extensions like cross-classified random effects. (Olson et al., 2020; Brunton-Smith et al., 2016).

challenging, especially for in-person assessments. In fact, there are two other recent papers that have studied enumerator effects in surveys through frameworks that approach a fully-interpenetrated designs in developing contexts. However, neither of these studies managed to reach the ideal level of individual-level randomization due to either the infeasible amounts of travel for enumerators this would have entailed (Di Maio and Fiala, 2018) or the logistical difficulty of enforcing individual-level assignments (Laajaj and Macours, 2017). In the case of Di Maio and Fiala's (2018) study, the authors were only able to randomly assign in-person enumerators to small geographical areas in Uganda within which it was still feasible for enumerators to travel. For the Laajaj and Macours (2017) study, the authors randomly assign enumerators to subjects at the individual level, but their compliance rate is only 75%. Because of similar constraints in educational assessment, assessors are rarely randomly assigned to respondents at the individual-level. As such the quantification of enumerator effects has been an elusive subject in the education literature.

Our paper offers two concrete contributions to the literature on enumerator effects, and their potential implications for educational assessments and impact evaluations. First, we document substantial "enumerator effects", or variation in how different assessors graded similar levels of performance, on a phone-based mathematics test, presenting some of the first evidence of enumerator effects on educational assessments. Through a fully-interpenetrated design in the administration of the assessment, we show that enumerator assignments explain 12 percent of the variation in the numeracy scores recorded. When we examine the survey questions, accounting for the enumerator assignment explains an even larger share of the variation in these responses than in the numeracy assessment. For example, enumerator assignment explains 32 percent of the variance in the likelihood of reporting a COVID-19-related income shock. We provide some evidence that younger teachers, teachers with fewer years at our partner's schools, teachers in charge of higher grades, and teachers at the same school as the child that they are assessing, record systematically higher math scores. This is suggestive evidence that enumerator effects, in this case, can be at least partially explained by observable characteristics of the assessors and the match along observables with their students.

Our second contribution is the quantification of the extent to which enumerator effects could yield spurious results in the context of a typical impact evaluation where enumerators are assigned to classes or schools rather than randomly assigned to individual students due to logistical considerations. Through simulation work, we find that assigning enumerators to whole classes or whole schools would yield point estimates that are statistically different from those found in the original impact evaluation about 10 and 13 percent of the time, respectively. This is in contrast to what we find when we simulate enumerator assignment at the level of individual students, where we obtain statistically different results from those in the original impact evaluation less than 1 percent of the time.

Following Evans and Yuan (2020), we estimate that had enumerators been assigned to whole schools for the companion RCT, there would have been more than a 1 in 10 chance of observing a treatment effect that was larger than the mean effect size in math, simply because of enumerator assignment at the school level. As a result, we ultimately recommend randomly assigning individual students to assessors, whenever feasible, to minimize bias. Our study also points to one advantage of phone-based assessments which is the relative ease with which assessors can be randomly assigned and bias therefore reduced compared to assessments administered in-person, particularly in the context of impact evaluation.

II. Study design

Context

This project was conducted in partnership with the organization NewGlobe, which operates as a technical partner of government-led education programs and supports its own community schools in several LMIC. One of these networks of community schools is Bridge Kenya, which is the context for our study. Our sample covered students across 105 private schools in 29 of the 47 Kenyan counties, and in all eight of the areas previously considered "provinces." There is a wide range in the socioeconomic characteristics of the locations covered. The local multidimensional poverty rate (at a 5 km radius from the school) ranges from 7.8% at the 10th percentile in our sample, to 59.2% at the 90th percentile. Although students in these schools and their families are relatively disadvantaged on a global scale, they are likely more socioeconomically advantaged and urban than typical families enrolled in Kenya's public schools. For example, nationally, 27 percent of families report the mother having no formal education while this is true for only one percent of our sample (Twaweza, 2014). Similarly, we estimate that the adult female literacy rate in the communities where pupils in our sample reside is 85%, compared to a national average of 79% (Bosco et al., 2017). The average student in our sample is 11.5 years old, as our sample consists of students in grades 3, 5, and 6. To be part of our sample, students needed to have access to a cellphone. However, this was not a very restrictive condition in this context, as the World Bank reports that there are 1.14 mobile subscriptions per person in Kenya. Finally, we show some additional descriptive statistics of our sample in Table 1.

The current study was part of a larger project conducted in Kenya, which comprised three different studies, all of which rely on overlapping sources of data. The first study consists of a randomized impact evaluation of individualized remote math instruction that teachers delivered by phone while schools were closed due to COVID-19 (Schueler and Rodriguez-Segura, 2021). The second project consists of the validation of phone-based assessments for early literacy and numeracy, and the exploration of the best uses for these

assessments based on their psychometric properties. That second study concludes that the phone-based assessments administered as part of this project demonstrated evidence of validity when used to measure aggregate performance, such as in the context of comparing a treatment and control group, but less evidence of validity for accurately tracking individual student-level performance (Rodriguez-Segura and Schueler, 2022). Finally, the third project consists of this study, which explores the presence and potential implications of enumerator effects in educational assessments. The impact evaluation study was pre-registered as AEARCTR-0006954 while the second study and current study were pre-registered in a separate pre-analysis plan related to educational measurement (AEARCTR-0006913).

Data collection and sampling

The collection of the phone-based assessment (PBA) data happened over 18 days in December of 2020 while face-to-face learning was on hold, after 9 months of school closures. These data were intended as one interim measure of learning outcomes for the impact evaluation of a remote instructional program via mobile phones and were needed because, at the time, it was unclear when students would return to school and take in-person assessments. The sample includes students in grades 3, 5, and 6 across all 105 schools. Due to budget constraints and response rate projections based on an earlier pilot, we selected a simple random sub-sample of students to be assessed from baseline performance blocks of students from all schools (6,295 students out of 8,319 in the impact evaluation sample were selected to be assessed). Of the 6,295 students on call lists, 2,644 were ultimately reached and assessed. Given the pace of enumerators and the rate at which successful assessments were completed, our initial target of 6,295 students proved too ambitions given the time constraints to collect the data. The sampling strategy and logistics are discussed in more detail by Rodriguez-Segura and Schueler (2022).

Assessors and assignment to students

Each of the 6,295 students on the call list was randomly assigned to one of 20 assessors. Randomization was accomplished at the student, rather than the class, school, or region level. Again, this allows us to attribute any differences in average scores across assessors to the assessors themselves rather than to differences in the performance or demographic characteristics of the different groups of students assessed by each enumerator. Among the 2,644 students that were ultimately assessed, the compliance rate to this random assignment was 98.8 percent. All results are robust to analyzing from the perspective of "assigned" assessors (in spirit, akin to an "intent-to-treat" analysis), or from the perspective of the "actual" assessor (similar in spirit to a treatment-on-the-treated analysis). The order in which assessors were asked to call students was also random. Strong protocols were in place to preserve this order, as assessors were centrally and simultaneously trained and had

continuous guidance for the first few days of data collection. The only information assessors had at the time of calling was the student's name, grade, and school, but no information about their previous performance or socioeconomic characteristics that they could have used to selectively call students.

All 20 assessors who worked on the data collection of the PBA were teachers within the Bridge Kenya system, for whom our partner had data on the grade and school where they taught and their years of experience. Our partner recruited these assessors to work full-time between December 7th and 23rd for data collection. Although assessors were full time Bridge teachers, typically PBA assessors did not know the students who they called to assess prior to the PBA administration. Only for 10 students (fewer than 0.4 percent) did the randomly assigned assessor end up being the student's own in-person teacher from the first two months of 2020. Assessors were paid by the day worked, and absenteeism was low. The minimum number of days worked for this wave were 9, the median 14, and the maximum 18. Assessors were aware that the PBA was a low-stakes assessment for the purpose of monitoring learning.

Numeracy assessment and survey questions

The phone-based assessment measured numeracy skills using 14 questions for students. These questions were divided into two sub-sections: the first 9 questions were part of the "core numeracy" sub-section, which asked the same questions to students from all three grades. This section included questions that ranged from counting, to basic operations, and finally a word problem. The second set of 5 questions were part of the "curriculum-aligned" sub-section, and these questions differed by grade. More specifically, these questions were designed so that they would assess concepts students would have been learning in class had in-person school been open.

The grading of each sub-section and of the assessment as a whole was done using a two-parameter item-response model, but we obtain almost identical results if we grade the exam through a simple percentage of the share of correct answers. All scores are standardized at the grade-level such that the mean for each grade is 0, and the standard deviation is 1. Rodriguez-Segura and Schueler (2022) provide a more rigorous psychometric exploration of the properties of this assessment.

After enumerators asked all 14 numeracy questions to students, they were instructed to also ask five survey questions to the parents or guardians of the children. These questions ranged in their degree of sensitivity: the less sensitive questions asked about the child's study habits and practices during school closures and these were asked first. The more sensitive questions asking about parental education ("*What is the highest level of school that you or someone in your household has completed?*", with eight discreet answer choices ranging from "Some primary school" to "Post-graduate degree", and COVID-19-related shocks to

the household's living conditions were left for the end ("*Finally, this has been a hard time for many families due to the coronavirus pandemic. In order to understand how this disruption has influenced children's learning outcomes, it would be helpful to know whether you have experienced any of the following since March when schools were closed*", with three potential answers including moving to a different home, health shocks, and changes to the household income). Conditional on a child starting the assessment, all children completed the assessment and all parents answered all five of the survey questions. The full assessment and survey are provided in Appendix A.

Administrative data and school information

We complement the phone-based assessment data with student-level administrative data provided by our partner. This data includes, for each pupil, individual-level covariates which include the gender, and age of the student, and school that they were enrolled at as of two months before the assessment. Similarly, we have access to three rounds of in-person baseline test scores in math, English, and Kiswahili collected before the phone-based assessment was administered and before school closures. These scores come from standardized tests administered across all 105 schools, in which, for any round of assessments, all students in a given grade took the same test. These data help us get a reliable and comparable measure of baseline achievement for all pupils.

School-level administrative data consists of the school's latitude and longitude, the total student enrollment, the pupil-teacher ratio in the overall student body and in the target grades (i.e., 3, 5, and 6), the female-male ratio in the overall student body and in the target grades, and the average principal, teacher, and student attendance rates. To learn more about the communities where these schools are located, we complement the administrative data with geospatial data containing information on community-level covariates. In particular, we use the GIS poverty rate raster layer from Tatem et al (2013), and the GIS adult (15-49) female literacy rate raster layer from Bosco et al (2017). We use each school's latitude and longitude of each school to create an average poverty rate and average adult female literacy rate for the 5-km circular area surrounding each school.

III. Methods

Broadly speaking, we are first interested in understanding whether there were systematic differences between the way individual assessors scored PBAs, including both the academic assessment and the survey responses. If so, we also want to understand whether observable characteristics of enumerators or the match along observable characteristics between enumerators and pupils—at least those for which we have data—predict higher or lower scores. Similarly, if we do observe enumerator effects, we want to quantify the extent to which this could become a problematic feature in the estimation of

policy-relevant statistics like treatment effects. Below we outline our methodological plan for answering these three main research questions.

How large were the enumerator effects in this assessment?

To explore whether enumerators recorded scores for similarly-performing students differently, in the past, the logistical challenges and research concerns of studying enumerator effects have typically been addressed by assigning in-person assessors to small geographic areas where it is still feasible for assessors to move around (Lupu and Michelitch, 2018). Although this is typically the best choice under realistic logistical constraints for in-person measurement tools, it might still allow for a high rate of "unlucky draws" in the assignment of enumerators to subjects, which when coupled with heterogenous practices across enumerators in data recording, might lead to spurious estimates simply because of the assignment of enumerators to clusters of observations. To avoid this issue, researchers have proposed individual-level assignment of units to assessors, "fully-interpenetrated" designs (West and Blom, 2017), when feasible. In our study, the full randomization of enumerators to students and high compliance rates with these assignments allow us to explore the extent to which enumerator effects may be present in PBAs. In other words, we are able to isolate the effect of enumerators from differences in achievement or other characteristics among the students assessed by different enumerators. Randomization allows us to assume that the achievement levels are the same, on average, across all groups of students assessed by different enumerators.

Our first approach to exploring the presence of enumerator effects is to follow Di Maio and Fiala (2018), Himelein (2016), and Laajaj and Macour (2017) by running a multivariate linear regression for each of the outcomes of interest (i.e., assessment scores) on enumerator fixed effects as the independent variables. We include fixed effects for each of the enumerators (19 in total, as one is the reference group). From this regression, we obtain a set of R^2 statistics which display the extent to which simply accounting for each student's enumerator explains variation in the outcome. Since the assignment of enumerators to students was random, one would expect this R^2 statistics to be close to zero if enumerators recorded scores in the same manner – in the same manner that enumerator assignment does not explain variation in our covariates (Table 1). Therefore, the higher the R^2 statistics, the stronger the evidence for the presence of enumerator effects.

The fully interpenetrated design also allows us to make predictions about the number of enumerators who are outliers in terms of the scores they recorded – that is, how many assessors recorded scores that were significantly different from everyone else's scores. In particular, given the individual-level assignment of assessors to students, we expect that the average score recorded by all assessors should be statistically equivalent for a large share for the assessors, in the absence of enumerator effects. For instance, we expect two of the 20 numeracy assessors to be statistically different from the rest at a confidence level of 90

percent. To address this, we test the extent to which each of our assessors is different from the rest, separately for numeracy and literacy. Specifically, for student i, and assessor m:

$$\text{Total score}_{ij} = \beta_0 + \beta_1 \text{Assessor}_m + e_{im}$$

Where Assessor_m is an indicator variable which takes the value of 1 if the assessor is assessor m, and 0 for all other assessors. This is repeated for each of the 20 assessors. The sets of β_1, with their respective confidence intervals, are recorded. Following, Von Hippel et al. (2016) and Von Hippel and Bellows (2018), we also use a Bonferroni correction for these confidence intervals to account for multiple hypothesis testing and to generate a null distribution against which we can compare the Bonferroni-corrected confidence intervals[3] As a robustness check, we repeat this exercise also controlling for the baseline score of each pupil, their grade, and their school. In other words, even if after randomization of assessors, an assessor obtained an imbalanced draw along these characteristics, this robustness check would account for this.

Did observable enumerator characteristics predict higher scores?

If we find evidence for the presence of enumerator effects, we also want to understand whether there are any observable characteristics of enumerators correlated with differential scores. This is particularly relevant as it could shed light on whether program managers could know a priori who among their enumerators might yield systematic differences, and hence select or train them accordingly. To tackle this question, we regress learning outcomes on different assessor characteristics, one covariate at a time, to test whether any of these characteristics predicts higher scores. We also leverage the fact that the allocation of students and enumerators was random, so that the match on their observable characteristics was also random. Hence, we can also explore the extent to which the match of student and teachers based on baseline student characteristics drives differential results. Specifically, for student i, assessor m, and observable X:

$$\text{Outcome}_{im} = \beta_0 + \beta_1(\text{Assessor and student match on X})_{im} + e_{im}$$

Where the variable "Assessor and student match on X" takes the value of 1 if the assessor and student share the same characteristic, and 0 otherwise. These characteristics include

[3] We follow this analytic approach while realizing that using Bonferroni-corrected confidence intervals and these null distributions are the most conservative approach to studying enumerator effects in this manner. In other words, if we detect enumerator effects through this approach, we would certainly detect them with looser approaches like tallying the share of instances in which β_1 displayed statistical significance, and checking if this share is lower than what one would expect by sheer chance, or using a null distribution that is 0 for all assessors.

whether the assessor and the student are based in the same school, whether they are assigned to the same grade, or whether the teacher is assigned to a similar age group as the child's grade (i.e., lower or upper primary). Taken together, the results from this section can inform whether enumerator effects, if at all, can be reduced by targeting specific sub-groups of enumerators of instances of assessor-student matches.

Could enumerator effects bias point estimates from an impact evaluation?

We also seek to understand the extent to which scoring differences across enumerators could bias the estimation of metrics like treatment effects in the context of impact evaluation. In particular, we ask whether different realizations in the allocation of enumerators to individual or organizational units would yield significantly different results solely due to differences in how enumerators record scores. In particular, we explore whether these results would vary were the enumerators allocated at the class- or school-level, as this has historically been the more common approach to assigning assessors for in-person assessments. To do so, we leverage the main intent-to-treat (ITT) estimates of the effect of phone-based tutoring on PBA numeracy scores from the field experiment that prompted the collection of the PBA data analyzed here (Schueler and Rodriguez-Segura, 2021). For the field experiment, treatment was assigned at the school-level, and there were two different treatment arms (T1 and T2). Schueler and Rodriguez-Segura (2021) find average ITT effects on the numeracy PBA scores of non-statistically significant β_1= 0.04 for T1 and β_1= -0.03 for T2, both in standard deviation units.

To explore the sensitivity of these point estimates to enumerator effects, we first predict the counterfactual PBA scores that students would have received under different enumerator assignments. To do so, we start by running the following model for student i, in grade j, at school k, assessed by enumerator m – who was the enumerator that actually assessed child i in the context of this study:

$$Y_{ijkm} = \beta_0 + \beta_1(\text{Baseline score})_{jjk} + \lambda_j + \mu_k + \eta_m + e_{ijkm}$$

Y_{ijkm} is the observed math score outcome, "Baseline score" represents the baseline in-person score for each student, and λ_j, μ_k, and η_m represent grade-, school-, and enumerator-fixed effects. This "calibrated model" yields a set of coefficients, all of which were estimates based on observed data, for each of these predictors. If one were to plug in a given student's covariates into this model and add the error term, one would obtain their actual PBA score.

After running this model with the, we proceed to randomly re-assign enumerators at different levels of aggregation to simulate alternate assignment scenarios as if we were starting the project from scratch. In particular, we randomly re-assign enumerators at the level of the students, then at the level of classes, and finally at the level of schools. These

last two steps mirror alternative study designs where enumerators are not assigned in fully-interpenetrated designs, but rather are clustered within some natural organizational unit. The simulation at the student-level allows us to benchmark the outcomes obtained for these other two levels of enumerator assignment to students.

Using the calibrated model, meaning the coefficients obtained from the first model with observed data, we plug each student's actual baseline score, grade, and school, along with their newly simulated random assignment to an enumerator n, to create a predicted test score $\hat{Y}_{ijkn}$ for each student:

$$\hat{Y}_{ijkn} = \beta_0 + \beta_1(\text{Baseline score})_{ijk} + \lambda_j + \mu_k + \eta_n + e_{ijkn}$$

$\hat{Y}_{ijkn}$ represents the predicted score one would expect student i to have obtained, had they been assessed by enumerator n (using the fixed-effects obtained for each enumerator in the calibrated model) rather than their actual enumerator m. For this exercise, each student receives three different $\hat{Y}_{ijkn}$: one under a new assignment of assessors at the student-level, another at the class-level, and another at the school-level. We then generate the predicted ITT estimates of the treatment effect of remote tutoring under the different enumerator assignments using $\hat{Y}_{is}$ as the outcome. This model yields $\hat{\beta}_1$ and $\hat{\beta}_2$, which are the simulated treatment effects had this specific realization of enumerator assignment been the actual assignment. We can then test whether the $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated with simulated enumerator assignments at the student-, class-, and school-level are statistically different from β_1 and β_2, that is – from the "true" ITT estimates observed in the field experimental study.

Finally, we repeat this simulation exercise 10,000 times for each of three levels of aggregation at which we assign assessors. This repeated exercise yields distributions of simulated treatment effects under the three levels of assignment of enumerators. Using each of these distributions, we tally the number of times that $\hat{\beta}_X$ was statistically different from β_x (in other words, the number of times the simulated treatment effect was different than the actual treatment effect). This number can be interpreted as the share of the time that we would have obtained spurious results when estimating a treatment effect simply as a result of the specific assignment of enumerators at different levels of aggregation.

IV. Results

Enumerator effects were present and non-trivial in size for both academic and survey measures

We find that the scores and responses that individual enumerators recorded did depend on enumerator assignment, and that size of these enumerator effects was non-trivial. Specifically, accounting solely for the assessor conducting the PBAs through a set of assessor fixed-effects explains 12 percent of the variance in the academic assessment score, (8

percent for core numeracy and 13 percent for curriculum-aligned scores). The enumerator effects were even larger when it came to the survey measures, where 32 percent of the variance for the likelihood of reporting of a COVID-19 related income shock, and 23 percent of the variance for self-reported parental education, were explained by differences in enumerator alone. To put these numbers into perspective, we can compare them against the findings of Di Maio and Fiala (2018) and Laajaj and Macours (2017), the two other papers in a developing context that also used some degree of randomization to quantify enumerator effects in surveys. The largest result presented by Laajaj and Macours is that enumerator assignment explains 9 percent of their variance in non-cognitive scores. Similarly, most of the results presented by Di Maio and Fiala (2018) report that enumerator assignment explains less than 5 percent of the variation for most questions, while their main finding is that for political questions, enumerator assignment explains around 30 percent of the variation in responses. Therefore, the enumerator effects that we find for the math scores on the PBA could be considered "medium-sized", while the magnitude of enumerator effects for the survey questions on the PBA are as large as Di Maio and Fiala (2018)'s largest finding – namely, politically sensitive questions in Uganda.

Another approach that we take to document the extent of enumerator effects in our data is to tally the number of assessors who report scores that are significantly different from a null distribution if there were no enumerator effects. We visually display these results for numeracy, literacy, and fluency in Figure 1. Through this approach, we find that at the 95 percent level of significance, five (25 percent) of the numeracy assessors recorded average scores that were significantly different from what one would expect to happen by chance if enumerators had no effect on the responses. This finding remains identical when we control for baseline performance, and include grade- and school-level fixed effects in the model above to account for potential unintentional differences in the random allocation of enumerators to students. Furthermore, as a placebo test, we also test the extent this is caused by differences in baseline performance, by running the same model as above but with the baseline score in each subject as the outcome. In this case, no assessor is statistically different from the null distribution. We observe even larger effects for the survey questions. For instance, 11 (65 percent), and 16 (80 percent) of the assessors are different from the rest when reporting a COVID-19 income shock, and for self-reported parental education respectively. In sum, we find very strong evidence for the presence of "enumerator effects", or the systematically heterogenous recording of similar answers by different enumerators for phone-based assessments of foundational numeracy skills and survey-based measures of parent-reported student learning time, parental education, and economic disruptions.

Observable enumerator characteristics did predict some differences in academic and survey responses

Since we do find systematic differences in how enumerators record scores, we explore which assessor characteristics drive these effects, as this could inform enumerator selection or the extent to which enumerators could be provided with targeted reinforcement before the data collection to reduce bias. It could also provide suggestive evidence regarding the mechanisms through which enumerator effects operate. We show results in Table 2. In general, assessors who taught higher grades during the school year recorded slightly higher scores, on average. Interestingly, the total number of students or the total number of students assessed on the same day do not predict differential assessment, which suggests that differential assessment is not driven by differential assessor pace.

A striking feature of Table 2 is that enumerator experience, measured in terms of days worked on the PBA data collection by the time a student was assessed, is correlated with higher recorded scores. Keep in mind, the order of calls was randomized. Although preserving the randomized order was not feasible 100 percent of the time, the actual order in which students were reached explains as little variance as the order assigned (which was randomized). To put the magnitude of this effect in perspective, the median number of days an assessor worked was 14. Between the first day, and the time enumerators were half-way done (day 7), the mean score had increased by 0.35 standard deviations. In fact, the mean over-estimation of in-class percentiles by PBA percentiles increased about 1.5 percentile every day. When the distribution by day worked is plotted, as in Figure 2, it is clear that there is a positive trend for the median, but the sharp increase in the mean is driven by a decrease in the lower tails as low scores become less frequent by day worked. In fact, the standard deviation of the numeracy score shrunk about 0.17 or 15 percent by day 7. Following the empirical strategy from the section on enumerator effects, we run a regression of total numeracy scores on day assessed-fixed-effects, which yields an R^2 of 0.05. In other words, the day of work for each assessor in which students were called explains 5 percent of the variance in scores, which is particularly interesting given that both of these assessment features were randomized at the individual level. In all, we find evidence that assessors did grade students differently over time, becoming more lenient toward the lowest-performing students the longer they were on the job.

Since the match along observables between assessors and students was also randomly determined, we study whether these matches predict differential scoring practices and display results in Table 3. In general, the results are not always consistent across specifications which may not suggest non-random sorting but rather may simply reflect the relatively small samples for which there is a match on these dimensions. However, there are two results that appear robust. First, teachers assessing students in the same grade as the grade they teach tend to record higher scores. It is unclear whether this is because the

teachers are more lenient, or the students perform better in these cases. Second, teachers who teach at the same school during the regular in-person school year as attended by the student being assessed, recorded lower levels of COVID economic disruptions, on average. Again, it is unclear whether this is because parents were less likely to disclose disruptions to teachers from their schools or enumerators were less likely to believe or record disruptions when reported by families attending the schools with which they were most familiar.

Enumerator effects can bias point estimates from an impact evaluation

Given the strong presence of enumerator effects, we also want to understand how effects of this magnitude could affect the estimation of treatment effects for a given impact evaluation, and the extent to which individual-level assignment of enumerators to students may reduce this bias on the estimands. To do so, we run the simulation described in the previous section, where we recreate the treatment effects that we see in a companion impact evaluation paper (Schueler and Rodriguez-Segura, 2021) under different simulated levels of enumerator assignments. We present the distributions of the simulated treatment effects at all three levels of disaggregation in Figure 3, along with a summary of these results in Table 4. Although the distributions of treatment effects at all three levels of disaggregation are centered around similar values in all three cases, the dispersion grows with the level of enumerator assignment. In other words, the probability of observing a large—either positive or negative—treatment effect solely as a result of the level at which the enumerator assignment occurred, increases significantly the larger the unit of aggregation assigned to each assessor. For instance, the 90th percentile of treatment effects for phone-based tutoring was 0.02 SD at the student-level, 0.07 SD at the class-level, and 0.12 SD at the school level.

To benchmark these scores, we use Evans and Yuan (2020), who claim that the average RCT in international education has a mean effect size in math of 0.09 SD, and that the effect size for math at the 60th percentile is 0.10 SD. In other words, had we assigned enumerators at the school level for the companion impact evaluation study, there was over a 1 in 10 chance of observing a treatment effect that was larger than the mean effect size for RCTs in similar contexts, simply because of the realized assignment of enumerators. In fact, we quantify in Table 4 the number of times that we observe significantly different results from the "true" ITT estimates. While student-level assignment of assessors yields similar results over 99 percent of the time, class- and school-level assignment of assessors yields different results between 10 and 13 percent of the time – at a rate 13 times higher than that at the student-level. In sum, the combination of enumerator effects and enumerator assignment at a clustered level might lead to potentially spurious evaluations of policies because of the random clustering of outcomes given the realized allocation of enumerators.

V. Discussion

In this paper, we present some of the first evidence from a fully interpenetrated study design documenting the presence of enumerator effects in educational assessments of academic achievement. We find significant differences, larger than what one would expect to arise by chance, in how numeracy scores were recorded across enumerators for similarly-achieving students. Consistent with previous research, we also document enumerator effects for parent survey measures that are even larger in magnitude than those observed for the academic tests. We find that the combination of enumerator effects of this magnitude, and a more aggregate level of assignment of assessors to students beyond the level of individuals, could yield spurious estimates of the estimands of interest, such as treatment effects in the context of an impact evaluation, at a worryingly high rate. In other words, we find evidence to support the claim that specific realizations of the assignment of enumerators to units can yield undesirable and heterogenous clustering of enumerators across treatment and control groups, which can in turn lead to differences in outcomes which are solely due to differences in how assessors recorded scores within their assigned clusters. While we explore the extent to which enumerator effects could bias treatment effect estimates, this phenomenon also has similar implications for a range of between-group comparisons beyond impact estimates.

In theory, the ideal solution to this issue would be the individual-level random assignment of enumerator to subjects, which in the case of phone-based assessments does not tend to pose significant logistical constraints. We show that this practice would, in most cases, diminish the extent to which the results obtained from a survey or assessment are biased due to enumerator effects by ensuring that any bias introduced by particular assessors is randomly distributed across the population of respondents and not systematically clustered within any group of respondents. Therefore, our study reveals this major potential benefit of phone-based assessments, which have gained some popularity in recent years in large part due to pandemic-induced disruptions to in-person schooling and assessment[4]. However, we also acknowledge that the exact model of enumerator randomization at the individual-level that we use in this paper is not always feasible when it comes to the collection of in-person field surveys and assessments. We propose several ideas to minimize bias from enumerator effects when randomization of enumerators to students is not possible. The first approach we propose is to make logistical efforts to assign enumerators to the smallest clusters that are feasible for them to assess. This may be insufficient to eliminate all bias, as in the case of the remote tutoring field experiment, we see that the biggest difference

[4] In another paper (Rodriguez-Segura and Schueler, 2022), we describe some of the potential and disadvantages of PBAs, like the significantly weaker correlations with baseline assessments compared to repeated in-person assessments. However, we also describe some of the advantages of PBAs, like being able to assess hard-to-reach populations at a fraction of the costs of in-person assessments, and the capability to use fully-interpenetrated designs.

in the rate of spurious effects due to enumerator effects is in the shift from student- to class-level assignment (the next most disaggregate unit after students), and not from classes to schools. That said, this may reduce bias in some cases. Another potential compromise would be to still randomly assign enumerators to students at the individual level, and group school visits within the same time window for an enumerator, if the travel distances between the natural clusters (e.g., schools) are not too prohibitive and the number of clusters is not too large. Then, the researcher could account for the day that each student was assessed in their models (via methodological approaches like day fixed-effects), while still being able to claim that there was some degree of full interpenetration in the study design. However, unlike phone-based assessments, this approach would not allow for the randomization in the overall order in which all students within a single enumerator were assessed.

The second approach we propose is to assign each assessor to a roughly equal number of clusters being compared to each other. For example, in the case of a field experimental study, assign each assessor to an equal number of treatment and control clusters. Alternatively, if the goal is to compare private to public schools, evenly distribute these two types of schools across assessors. This approach would be a first step towards offsetting the differential scoring of assessors in the estimation of treatment effects or differences between other groups of interest. Similarly, this approach would allow researchers to include assessor-level fixed effects in their specifications, as it would provide common support across both sides of the treatment effect estimation for each assessor. The inclusion of assessor fixed effects would then de-mean the outcome from each assessor's idiosyncratic scoring bias. However, this approach is only feasible when there are fewer assessors than clusters. There is also the opposite scenario, where the clusters are so large that they require more than one assessor. In this case, the random allocation of assessors to the groups of assessors that visit different clusters seems particularly important to "dissolve" enumerator differences within clusters as much as possible.

Third, we find that, when it comes to academic assessments, whether or not the assessor teaches the grade of the student being assessed matters for the responses recorded. Therefore, assessment administrators would be wise to distribute these matches even across groups being compared. Relatedly, we observe that teachers from the same school as the family being surveyed record fewer economic shocks. This suggests that survey administrators should therefore carefully consider whether assessors are from the same community as the respondents when being tasked with asking sensitive questions, and again, attempt to spread these matches out across any groups being directly compared on the survey outcomes.

Fourth, there are potentially statistical adjustments that can be performed after data collection, and which could reduce some of the error introduced by the enumerator effects. For example, one could imagine a typical set up where enumerators are rigorously trained,

and then randomly assigned to large clusters like schools to collect reading fluency data from grade 3 students. At the end of the data collection process, they would all be shown the same pre-normed footage of several children with different performance levels taking the same fluency assessment that the enumerators just administered. The enumerators are individually asked to score each of the students in those videos, without knowing what the actual normed score of each video is. Then, each enumerator's scores are compared to the pre-normed scores for the videos to understand the magnitude and direction of the bias for each assessor (e.g., "enumerator A's tendency was to record scores that were on average 0.5 SD above a pre-normed score, while enumerator B's tendency was on average only 0.1 SD below the pre-normed scores). Finally, this information could be used for post-data collection adjustments to the observed scores. Questions along these lines, to the best of our knowledge, have remained largely unexplored in the literature, and as such, this is an area that is ripe for future research.

In terms of logistical considerations, we also recommend robust training of assessors. Training in general might have several goals, like reducing the extent to which enumerators make data entry errors, or maximizing the probability of subjects agreeing to participate in the interview. However, if the skills that are being trained are heterogeneously distributed across enumerators at the beginning of the training, pushing all these goals forward can also contribute to the reduction of enumerator effects. In this sense, there are valuable publicly available resources to incorporate best-practices into enumerator training (World Bank DIME, 2022). Having said that, researchers and practitioners may also find that there is only a certain amount of training that can sustainably take place given the financial and time constraints of a project, and as such, it is critical to incorporate into the training the most effective practices to successfully train enumerators.

The training of enumerators does not need to end at the beginning of the data collection process, and in fact, we believe that constant "norming" of assessors throughout the life cycle of the process can be beneficial to reduce enumerator effects. In particular, in this study we also find that assessors' behavior changes over time – which in this case manifested as fewer low numeracy scores recorded in the later stages of the data collection process suggesting either increased leniency toward lower achieving students by enumerators over time or a greater ability to detect achievement among students typically perceived as low performing with greater experience. Regardless of the mechanism, to avoid this phenomenon, other literatures have considered the practice of "norming" (Cohen and Goldhaber, 2016), that is, the conduction of frequent exercises that align enumerators amongst themselves, and with an ideal scoring behavior over time. Additionally, future research should explore the extent to which norming can minimize enumerator effects.

We also see implications for researchers drawing on previously published studies or datasets that rely on one-on-one assessments for which assessors were assigned at high

levels of aggregation, such as at the school, district, county, or country level. Firstly, such estimates should be interpreted with caution and may be over- or under-estimated due to the presence of enumerator effects. For areas of study in which this is the only method that has previously been used to gather outcome data, researchers should prioritize new studies that are not susceptible to this form of bias, if at all possible. Additionally, on topics for which there are puzzling discrepant results across studies, enumerator effects could be one explanation for these differences if there is variation in the level of enumerator assignment across studies. Secondly, for publicly available data sets, we recommend that the maintainers of these data also include certain de-identified information about the enumerators as part of the data set. For each observation, researchers should know, to the extent possible, who the enumerator who collected this data was. Then, this information should then be able to be linked to additional information about the enumerator that may potentially explain some of the enumerator effects that could manifest in this data set, if at all. For instance, for a data sets on early-grade reading fluency, understanding each enumerator's educational background and experience collecting early-grade literacy data would be a valuable addition to the data set.

Our hope is that this study is a starting point for future research on the implications of enumerator effects in other types of assessments, subjects, modes of assessments, and contexts. For instance, one question that remains from our study is the extent to which we observed enumerator effects in this case because the assessment happened over the phone. Paradoxically, it was the fact that the assessment was conducted over the phone what allowed us to isolate the enumerator effects in a clean empirical way. Future research should test whether these findings replicate with in person assessments and assessments of additional content areas such as literacy. Another important line of research is the extent to which different types of training and norming of assessors could reduce the extent to which enumerator effects can manifest in educational assessments. In other words, we do not know whether the enumerator effects that we observe here happened because of the training that they received, or in spite of it.

In all, the study of quantification of enumerator effects in educational assessments, particularly in LMIC, remains a nascent area of study. However, we present evidence making the case that researchers should pay closer attention to how these systematic differences between enumerators might be affecting their results, and ultimately, the conclusions that can be drawn from their studies.

VI. References

Adida, C.L., Feree, K.E., Posner, D.N., Robinson, A.L. (2016). Who's asking? Interviewer coethnicity effects in African survey data. Comparative Political Studies. 49: 1630–60

Baird, S., & Özler, B. (2012). Examining the reliability of self-reported data on school participation. Journal of Development Economics, 98(1), 89–93. https://doi.org/10.1016/j.jdeveco.2011.05.006

Benstead, L.J. (2014). Does interviewer religious dress affect survey responses? Evidence from Morocco. Politics and Religion 7: 734–60

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K., & Sheridan, M. (2016). Measuring the measurement error: A method to qualitatively validate survey data. Journal of Development Economics, 120, 99–112. https://doi.org/10.1016/j.jdeveco.2016.01.005

Blaydes, L., Gillum, R.M. (2013). Religiosity-of-interviewer effects: assessing the impact of veiled enumerators on survey response in Egypt. Politics and Religion 6: 459–82

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location-scale model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(2), 551–568. https://doi.org/10.1111/rssa.12205

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. Educational Researcher, 45(6), 378–387. https://doi.org/10.3102/0013189X16659442

Di Maio, M., & Fiala, N. (2020). Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect. The World Bank Economic Review, 34(3), 654–669. https://doi.org/10.1093/wber/lhy024

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. International Journal of Educational Development, 40, 315–322. https://doi.org/10.1016/j.ijedudev.2014.11.004

Evans, D. K., & Mendez Acosta, A. (2021). Education in Africa: What Are We Learning? Journal of African Economies, 30(1), 13–54. https://doi.org/10.1093/jae/ejaa009

Evans, D., & Yuan, F. (2020). How big are effect sizes in international education studies? CGD Working Paper 545.

Lupu, N., & Michelitch, K. (2018). Advances in Survey Methods for the Developing World. Annual Review of Political Science, 21(1), 195–214. https://doi.org/10.1146/annurev-polisci-052115-021432

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (2020). The Past, Present, and Future of Research on Interviewer Effects. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), Interviewer Effects from a Total Survey Error Perspective (1st ed., pp. 3–16). Chapman and Hall/CRC. https://doi.org/10.1201/9781003020219-2

Randall, S., Coast, E., Compaore, N., & Antoine, P. (2013). The power of the interviewer: A qualitative perspective on African survey data collection. Demographic Research, 28, 763–792. https://doi.org/10.4054/DemRes.2013.28.27

Rodriguez-Segura, D., Schueler, B. E., (2022). Can learning be measured by phone? Evidence from Kenya. EdWorkingPaper: 22-517. Retrieved from Annenberg Institute at Brown University: https://- doi.org/10.26300/gc6v-qv41

Schueler, B. E., Rodriguez-Segura, D. (2021). A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya. EdWorkingPaper: 21-432. Retrieved from Annenberg Institute at Brown University: https://- doi.org/10.26300/43qs-cg37

Strauss, M. E., & Smith, G. T. (2009). Construct Validity: Advances in Theory and Methodology. Annual Review of Clinical Psychology, 5(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Twaweza. (2014). Uwezo data-Household data https://www.twaweza.org/go/uwezodatasets

Varly, Pierre. (2020). Learning assessments in Sub-Saharan Africa. SDG4-Education 2030 in Sub-Saharan Africa. Analytic Report N°1. https://learningportal.iiep.unesco.org/en/library/learning-assessments-in-sub-saharan-africa

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. Economics of Education Review, 64, 298–312. https://doi.org/10.1016/j.econedurev.2018.01.005

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable?

Which programs are different? Economics of Education Review, 53, 31–45. https://doi.org/10.1016/j.econedurev.2016.05.002

West, B.T., Blom, A.G. (2017). Explaining interviewer effects: a research synthesis. Journal of Survey Statistics and Methodology. 5: 175–211

VII. Tables and figures

Table 1: sample description by baseline covariates

		Mean / (SD)	Share of variance explained by assessor assignment
Student characteristics	Student is female	0.50	0.3%
		(0.50)	
	Student's age	11.5	1.2%
		(1.7)	
	Latest standardized score math pre-school closures, by grade	0.02	1.0%
		(0.98)	
	Latest standardized score Kiswahili pre-school closures, by grade	0.02	0.9%
		(1.00)	
	Latest standardized score English pre-school closures, by grade	-0.00	0.7%
		(0.98)	
School characteristics	Total school enrollment	273.1	0.7%
		(73.0)	
	Attendance rate of school principal	0.89	0.4%
		(0.26)	
	Attendance rate of teachers	0.84	0.9%
		(0.15)	
	Attendance rate of pupils	0.51	0.7%
		(0.15)	
Community characteristics	Population within a 5km radius	261753.8	0.9%
		(376719.2)	
	Average female literacy within a 5km radius	0.85	0.9%
		(0.14)	
	Average poverty rate within a 5km radius	0.33	0.9%
		(0.17)	
	Distance to nearest cell tower (km)	0.42	0.6%
		(0.80)	
	Observations	2552	

Notes: the column displaying the share of the variance of each covariate that is explained by assessor assignment corresponds to the R^2 resulting from regressing each covariate on a set of fixed effects for each assessor. This follows the first methodological approach outlined in the "Methods" section.

Table 2: assessor predictors of outcomes collected through phone-based assessment

		Math score		Time reported studying		Reports using books		Parental education		Reports COVID-related income shock	
	Mean	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Years at Bridge as a teacher	6.25	-0.02***	-0.02**	0.07***	0.07***	-0.01**	-0.01**	-0.03***	-0.04***	0.03***	0.02***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Assessor grade taught	1.41	0.02***	0.03***	0.12***	0.12***	-0.05***	-0.05***	0.07***	0.07***	0.00	0.00
		(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Total students assessed	139.36	0.00***	0.00***	0.01***	0.01***	0.00***	0.00***	0.00	0.00	0***	0***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Number of students assessed by assessor on the same day	12.85	0.00	0.00	0.04***	0.04***	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Total days worked	14.04	0.00	0.00	-0.1***	-0.09***	0.01*	0.01*	-0.01	-0.01	0.04***	0.04***
		(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Enumerator experience (days)	7.26	0.05***	0.05***	-0.04***	-0.04***	0.01***	0.01***	0.01***	0.01***	0.02***	0.02***
		(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
School FEs, Grade FEs, baseline performance		N	Y	N	Y	N	Y	N	Y	N	Y
Outcome mean		-0.010		2.650		0.610		2.090		0.750	
Outcome SD		0.990		2.020		0.490		0.710		0.430	
Observations		2552	2552	2552	2552	2552	2552	2509	2509	2552	2552

Notes: Each coefficient comes from running a regression of the outcome on each assessor characteristic (predictors). * p<0.10, ** p<0.05. *** p<0.01.

Table 3: effect of matching characteristics between assessor and student characteristics

		Math score		Time reported studying		Reports using books		Parental education		Reports COVID-related income shock	
	Mean	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Same school	0.11	0.14**	0.06	-0.48***	0.71**	0.18***	-0.02	-0.04	-0.11	-0.28***	-0.22***
		(0.06)	(0.11)	(0.1)	(0.28)	(0.03)	(0.05)	(0.04)	(0.07)	(0.03)	(0.06)
Same grade	0.04	0.24**	0.19*	0.14	-0.15	-0.25***	0.00	0.43***	-0.03	-0.16***	-0.01
		(0.09)	(0.11)	(0.21)	(0.21)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)
Same age group of student, and class taught by assessor (lower primary vs. upper primary)	0.26	-0.02	-0.01	-0.31***	-0.29***	0.09***	-0.02	-0.07**	0.02	-0.04*	-0.04**
		(0.04)	(0.05)	(0.07)	(0.07)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
School FEs, Grade FEs, baseline performance		N	Y	N	Y	N	Y	N	Y	N	Y
Outcome mean		-0.010		2.650		0.610		2.090		0.750	
Outcome SD		0.990		2.020		0.490		0.710		0.430	
Observations		2552	2540	2552	2540	2552	2540	2509	2497	2552	2540

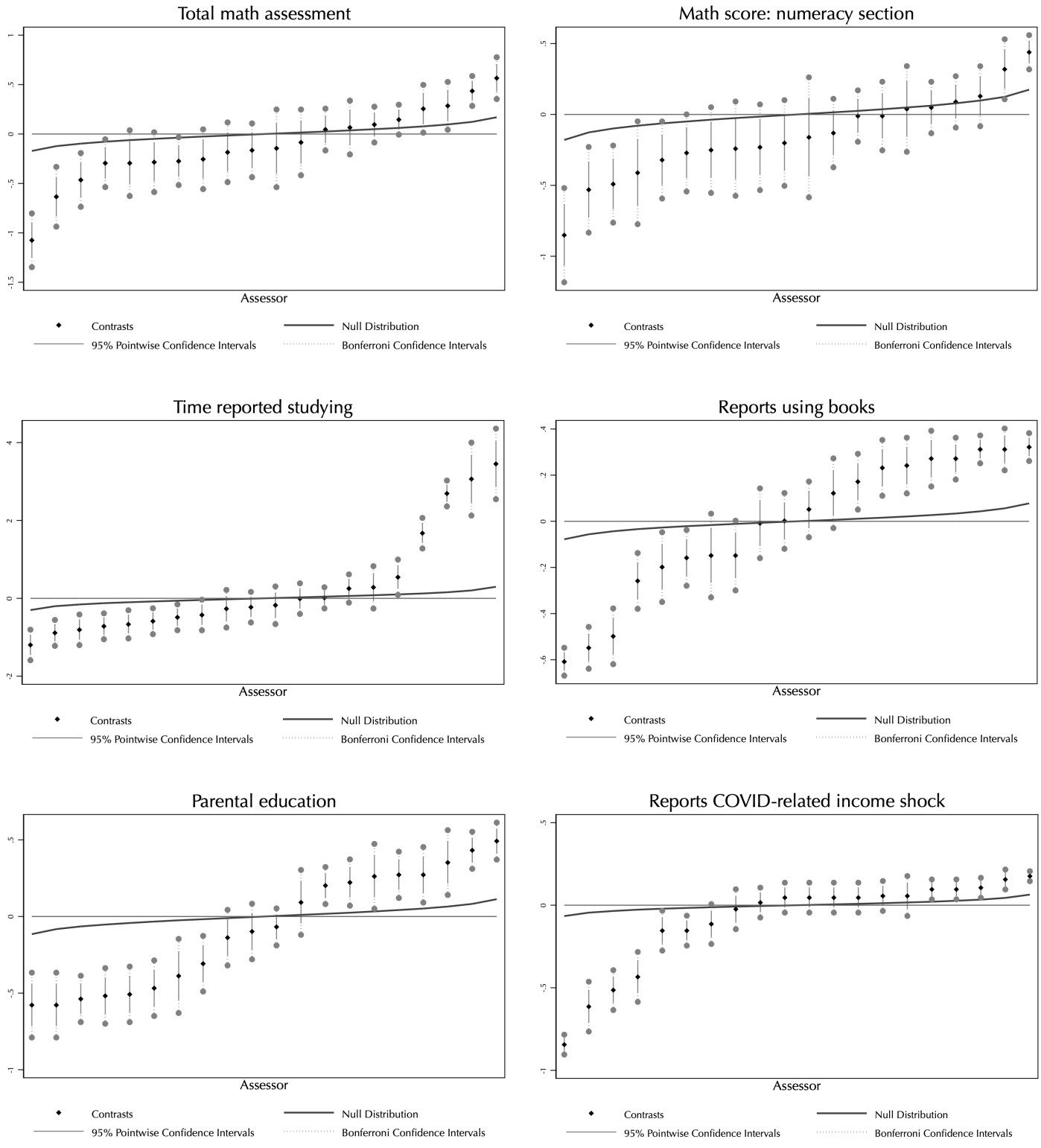
Notes: * p<0.10, ** p<0.05. *** p<0.01.

Table 4: percentage of simulation exercises that yielded treatment effects different from the observed treatment effects

Level of enumerator assignment	Treatment 1 (T1)	Treatment 2 (T2)	Average T1 and T2	Times higher than student-level
Student	1.4%	0.3%	0.9%	-
Class	11.0%	8.1%	9.6%	11.2x
School	13.0%	12.0%	12.5%	14.7x

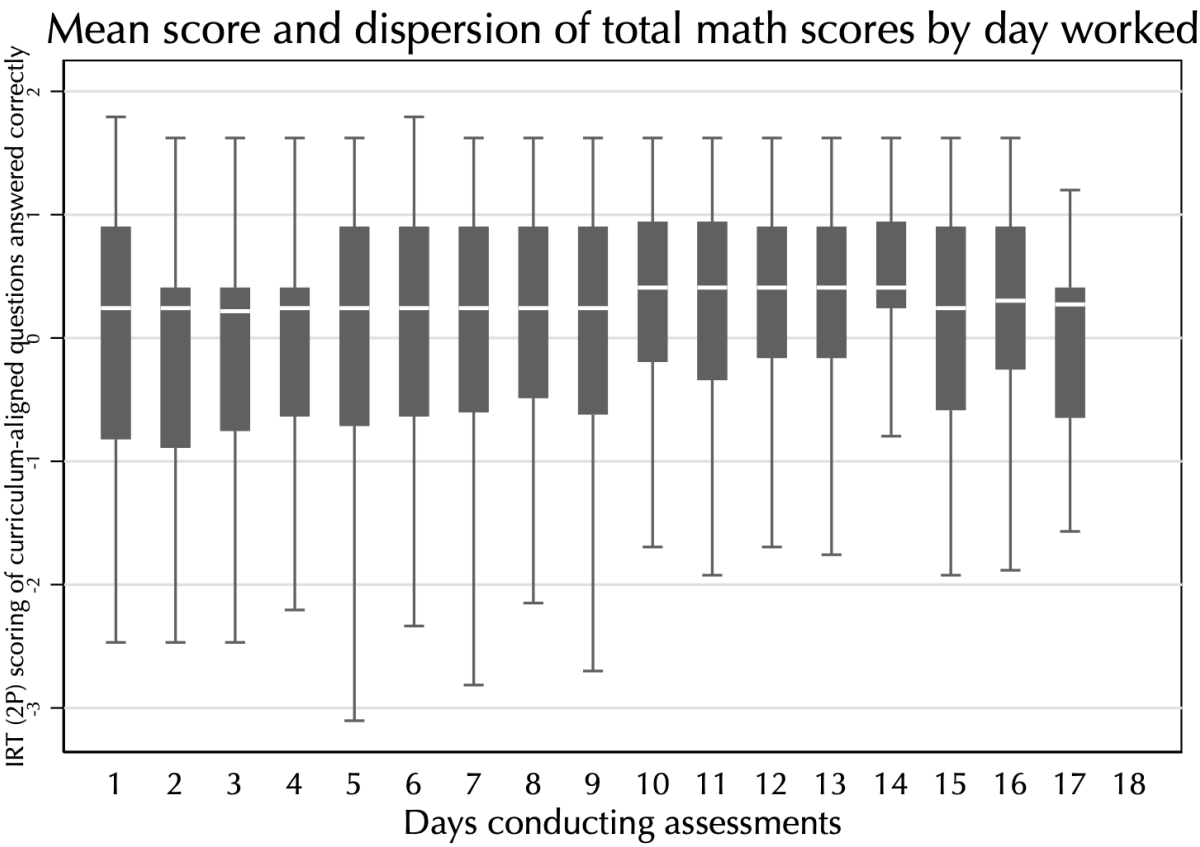
Notes: numbers obtained from simulating outcomes 10,000 times for each level of enumerator assignment

Figure 1: Differences between each assessor and all other assessors, by outcome



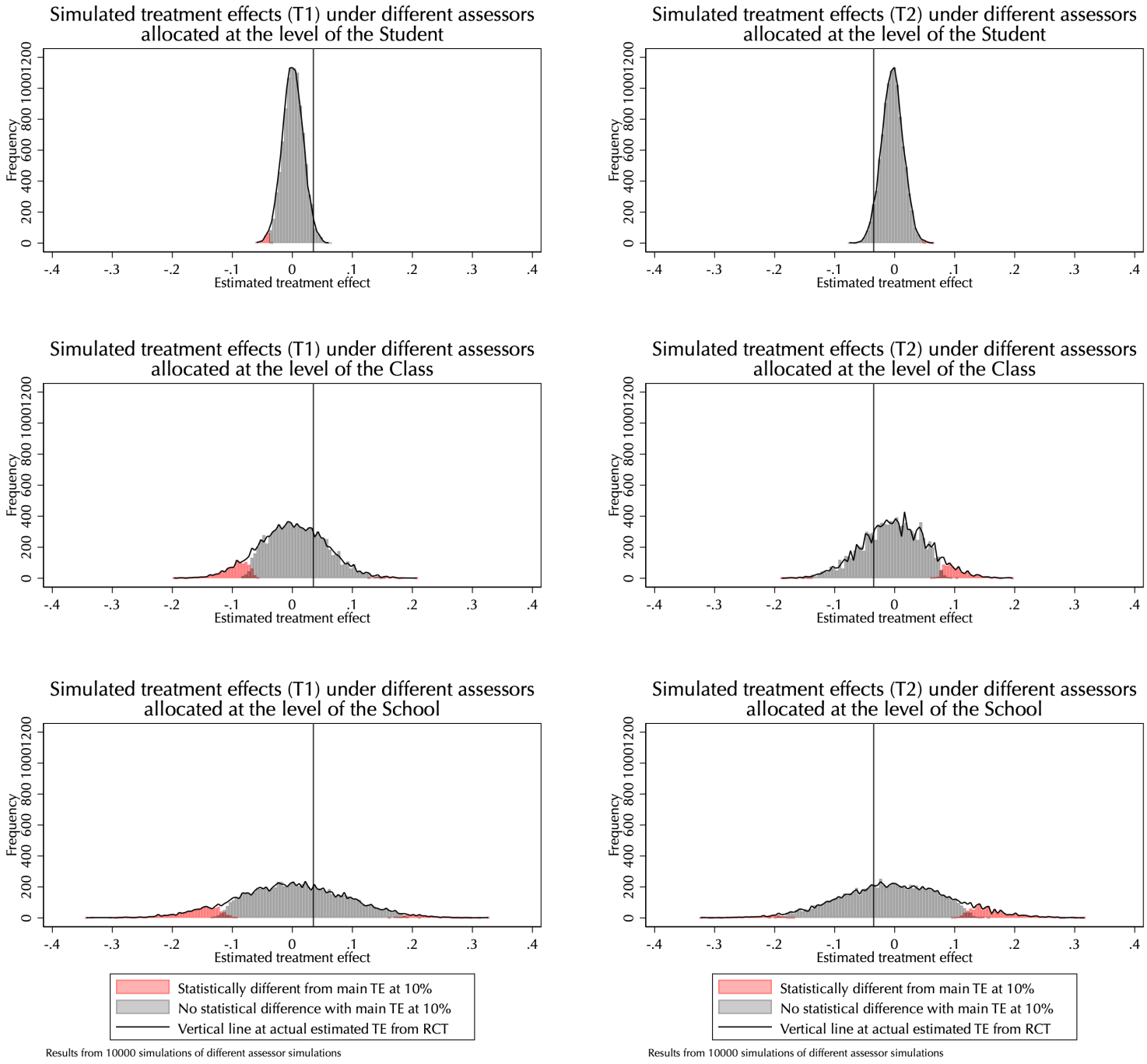
Notes: The difference between each enumerator and the rest is computed by regressing the outcome on an indicator variable, separately for each enumerator. All specifications include a grade- and school-level fixed effects. Coefficients sorted from left to right by plot, meaning that enumerator number does not necessarily match across panels. Standard error bars shown at the 95% level of significance. Standard errors are clustered at the school-level.

Figure 2: Distribution of total numeracy scores by the day worked for each assessor



Notes: Box plot showing the 25th, 50th, and 75th percentile of the total score attained on the phone-based assessment by the day that each child was assessed, according to the days that their assessor worked. Assessment graded using the first component of a principal component analysis.

Figure 3: Distribution of simulated treatment effects under different allocations of enumerators, by treatment arm and level of enumerator assignment



Appendix

a. Instrument used

Numeracy – Grades 3, 5, and 6

Section	Item number	Questions only for Grade 3	Questions only for Grade 5 and 6
Instructions		Assessor's name: _____ *[Anything written in this font is a note to assessor that should not be read aloud]* **Step 1: Introduction.** Hello. My name is _____ and I am calling on behalf of Bridge International Academies. I am hoping to speak first with your pupil __<insert name>___ about some math problems and then at the end to speak with you again to get a bit of information about how the term has been going for your family. Does that sound okay? **Step 2: Instructions.** First, I would like your child to work on a couple of maths problems. I ask that you put the phone on speaker or repeat out the questions to the pupil to answer. Please have your child answer the problems on their own on a scrap paper. After they are done with a problem, you or they can read out their answer to me. The answers will not count toward grades in school, so it's okay if your child does not get all of the answers correct. Is your pupil ready? **Step 3: Assessment.** ***Core Numeracy Questions*** *[Please ask students the following questions in order. If they get three questions in a row wrong, please do not ask any more of the "core numeracy questions" and move on to the "grade level questions" section below.]*	
Learning assessment – Core numeracy	1	Can you count from 20-30? *[Mark the highest number the student reached consecutively, so if they get to 27 but skipped 25, mark (e) 24]* o No answer o 23 o 27 o 20 o 24 o 28 o 21 o 25 o 29 o 22 o 26 o 30	
	2	Which is greater? 64 or 38? o Correct (64) o Incorrect o No answer	
	3	What is 62+18? o Correct (80) o Incorrect o No answer	
	4	What is 33+49? o Correct (82) o Incorrect o No answer	
	5	What is 43-20? o Correct (23) o Incorrect o No answer	
	6	What is 81-43? o Correct (38) o Incorrect o No answer	
	7	What is 3x4? o Correct (12) o Incorrect o No answer	
	8	What is the result of 8 divided by 2? o Correct (4) o Incorrect	

		○ No answer	
	9	Oil is 200 shillings per liter and rice is 100 shillings a kilogram. How much should I pay for 3 liters of oil and 4 kilograms of rice? ○ Correct (1000 shillings) ○ Incorrect ○ No answer	
Instructions		***Grade Level Questions*** *[Please ask students the following questions in order. If they get three questions in a row wrong, please do not ask any more of the "grade level questions" and move on to the survey question.]*	
Learning assessment – Curriculum-aligned items	10	Complete the following number pattern: 13, 19, ___, 31 ○ Correct (25) ○ Incorrect ○ No answer	What is 3487+2325? ○ Correct (5812) ○ Incorrect ○ No answer
	11	What is 145+213? ○ Correct (358) ○ Incorrect ○ No answer	What is 4756-2149? ○ Correct (2607) ○ Incorrect ○ No answer
	12	What is 278-124? ○ Correct (154) ○ Incorrect ○ No answer	What is 42x26? ○ Correct (1092) ○ Incorrect ○ No answer
	13	What is 8x5? ○ Correct (40) ○ Incorrect ○ No answer	What is the result of 96 divided by 12? ○ Correct (8) ○ Incorrect ○ No answer
	14	What is the result of 35 divided by 7? ○ Correct (5) ○ Incorrect ○ No answer	What is the result of 3/7 (three sevenths) + 1/4 (one fourth) ○ Correct (19/28 (nineteenth twenty-eighths) ○ Incorrect ○ No answer
Instructions		**Step 4: Student survey.** Nice work. Next I am going to ask you a general question about school. There is no right or wrong answer, please just give your best response.	
Survey – Students	15	How much do you feel your teacher cares about your learning during the remote learning period? *[Read out each answer choice]* ○ Not at all ○ A little bit ○ Some ○ Quite a bit ○ A lot	
Instructions		*[To child]:* Thank you very much. Now, I would like to ask your parent a few questions, could you put them back on? **Step 5: Parent survey.** *[To parent]:* Thank you. Now I would like to ask you a bit about your child and household during this period of school shutdowns. Your participation is totally voluntary and you are welcome to skip any questions that you do not feel comfortable answering.	
Survey – Parents	16	On average over the past week, how many hours a day has your child spent on education? *[This is in reference to the child who completed the test]*	
	17	How many times has your child's teacher called you or your child by phone in the past 7 weeks? ○ 0 ○ 4 ○ 8 ○ 1 ○ 5 ○ 9 ○ 2 ○ 6 ○ 10 or more ○ 3 ○ 7 ○	
	18	What are children in your household currently doing to learn?" *[Read out each option and mark all that apply]* ○ Educational TV programs or radio ○ Bridge@home ○ Receiving calls from child's teacher/academy manager/academy ○ Educational content on the internet	

		○ Books we have in the household ○ Government educational content - courses, audiobooks, or lessons ○ I/Others in my household are teaching or reading with them ○ I/Others encourage children to do distance learning (radio, television, phone, etc.) but do not help ourselves ○ We are paying for in-person tutoring ○ Other [*Please specify*]:_____ ○ Nothing
	19	What is the highest level of school that you or someone in your household has completed? ○ Some primary school ○ Primary school completion ○ Some secondary school ○ Secondary school completion ○ Certificate or other post-secondary ○ Some university ○ University completion ○ Post-graduate degree
	20	Finally, this has been a hard time for many families due to the coronavirus pandemic. In order to understand how this disruption has influenced children's learning outcomes, it would be helpful to know whether you have experienced any of the following since March when schools were closed: *[Read out options, pausing after each option for a yes/no, and mark all that apply]* ○ Moved to a different home ○ Had someone in your home experience health challenges ○ Had changes to your job or income
Instructions		**Step 6: Closing.** Many thanks for your help with this.