Pre-Analysis Plan
August 5th, 2020

I. PURPOSE

The inability to consistently deliver at large scale promising education interventions is an important contributing cause to inequality in the U.S. The research team applies insights from price theory and field-based randomized controlled trials to examine the effect of implementing a promising academic skills development program at large scale *before* implementing at scale. The project is designed to provide evidence of direct scientific and policy value for attempts to scale up a specific intervention, but also stimulate a much more thorough investigation of social policy scale-up challenges by refining these methods and demonstrating their feasibility and value.

The research team examines the challenge of program scale up for a promising intervention studied in Chicago at medium scale in the past – SAGA tutoring. Past work has demonstrated that SAGA's intensive, individualized, during-the-school-day math tutoring can generate very large gains in academic outcomes in a short period, even among students who are many years behind grade level. This study will explicitly explore the extent to which there is a trade-off between effectiveness and scale for this intervention. By taking advantage of the power of random sampling, this study will also allow for observation of the program's effectiveness as if it were running at three-and-a-half times the proposed scale in a subset of the study population.

II. RESEARCH QUESTIONS

The University of Chicago Education Lab and Crime Lab New York research teams are carrying out a randomized controlled trial during the 2016-17 and 2017-18 academic years to build on previous collaborations with the Chicago Public Schools (CPS), the New York City Department of Education, and SAGA Innovations that have found that SAGA's intensive, individualized, during-the-school-day tutoring can generate very large gains in academic outcomes in a short period of time, even among students who are many years behind grade level. This research suggests the promise of this approach for improving the academic skills and educational attainment of disadvantaged youth, even once they have reached adolescence. However, to truly affect outcomes at the local and national level, SAGA would have to be rolled out on a much greater scale than researchers have been able to study in Chicago. Yet little is known about how to take promising interventions to scale. This study seeks to build the science of scale-up, by examining the extent to which this individualized tutoring program can be implemented at an even greater scale and by explicitly exploring the trade-offs between effectiveness and scale.

The SAGA Innovations program expands on the nationally recognized innovation of high-dosage, in-school-day tutoring developed in Match Education's charter school in Boston. The tutoring program meets as a scheduled course, Math Lab, once a day during the normal school day, and is provided in addition to a student's regular math class. Students work two-on-one (two students with one tutor) with the same full-time, professional tutor for the entirety of the school year. The content of the tutoring sessions is aligned with what students are learning in their regular math courses, but is also targeted to address individual gaps in math knowledge. Also following the original model developed by Match Education, SAGA tutors use frequent internal formative assessments of student progress to individualize instruction.

A previous randomized controlled trial conducted by the University of Chicago research team found that one year of this intervention, delivered in AY2013-14 in the Chicago Public Schools, generated between one and two extra years of academic growth in math, over and above what the normal U.S. high school student learns in one year (Cook et al., 2015; Reardon, 2011). The estimated effects for math achievement are on the order of 0.19 to 0.30 standard deviations, depending on the exact test and norming used. The intervention also improved student grades in math by 0.58 points on a 1-4 grade point scale, compared to a control mean of 1.77. These gains are particularly important because of the growing evidence on the importance of math specifically for short- and medium-term success in school, and for long-term life outcomes such as employment and earnings (Duncan et al., 2007).

This study aims to build upon the investigators' previous evaluations of the program, and will provide insight into the ability of this program to serve youth at a much larger scale. Specifically, this study aims to answer the following research questions:

(1) What is the effect of implementing an evidence-based individualized tutoring program at larger scale?

(2) What is the relationship between the effect of the program and the scale at which the program is implemented?

III. STUDY DESIGN

Implementation sites are divided into two sets: sites in Chicago at which students are randomized to receive tutoring (hereby referred to as “scale-up” schools), and sites in Chicago and New York City where principals have primary discretion to choose which students receive tutoring (hereby referred to as “returning schools”). In addition to randomizing students into the SAGA program, scale-up schools are also served by randomly selected tutors. The research team is having SAGA over-recruit tutors as though they were implementing at larger than the intended scale in the scale-up schools. Investigators then randomly select one in three-and-a-half tutor applicants to continue through SAGA’s standard hiring process, and positions at the scale-up schools are only filled by these randomly selected tutors. All study schools (both scale-up and returning) are implementing a third form of randomization: students in the SAGA program are randomly assigned to tutors.

In order to study research question #1, investigators will take advantage of the power of random sampling of tutors and the random assignment into the SAGA program in the scale-up schools to study scale-up of this program without actually having to implement the program at a much larger scale. By comparing the outcomes of students randomly assigned to the SAGA program to students randomly assigned to the control group in these schools, we will be able to rigorously estimate the effects of the program if it were being staffed by the tutors that would work for a program that is three-and-a-half times as large as the program currently operating in the scale-up schools.

To gain insight into research question #2 above, which seeks to determine the relationship between program scale and effects, tutors at all sites are ranked by SAGA leadership based on relative expected quality. Because the research team is randomizing student pairs to tutors, we will be able to isolate the effect of value-add of each tutor in the SAGA program. Combining this information with the SAGA rankings of applicant quality, the researcher team will be able to examine tutor ranking and tutor effectiveness. Assuming that the program would hire tutors in the order of their ranking depending on the number of tutor slots they needed to fill, this analysis will shed light on the relationship between scale and effectiveness.

Lastly, from our experience in Chicago, we have learned that it is very important to have a deep understanding of program implementation. As such, our evaluation team will conduct program observations of tutoring sessions using a structured observation protocol that documents key aspects of program implementation (e.g. student and tutor engagement in the program). Research team members will visit each school at least once a month for most of the school year.

IV. STUDY SAMPLE

Our study will run for two academic years: academic year (AY) 2016-17 and academic year 2017-18. School sites will be divided into Chicago and New York. Only about half of the Chicago sites will randomize tutor applicants to be hired, and will randomize students into treatment. No New York City sites will undergo the tutor or student randomization. All sites across Chicago and New York City will undergo the student-tutor randomization.

The research team will work with SAGA, the Chicago Public Schools, and the New York City Department of Education to select schools that serve economically disadvantaged and minority students. SAGA will hire about 100 tutors each year in Chicago, and about 50 each year in New York City. As per SAGA's standard model, each tutor will work with approximately 14 students.

For tutor applicant randomization, tutor selection will be performed using tutor applications. Tutors who make it through SAGA's first screen of hiring – e.g. a resume screen – will be randomly selected via a one-in-three-and-a-half randomization rate.[1] Those who are randomly selected via this process are eligible to be interviewed by SAGA to be hired and work in the same schools where students are randomized to treatment. These tutors will be assigned to serve in the Chicago schools that are randomizing students into the SAGA program (the scale-up schools).

To conduct student randomization to treatment, we will use existing school administrative data. Eligible students will be randomly assigned to one of two study arms via a fair lottery—SAGA Innovations' high-dosage tutoring, or status quo (control). The research team will use administrative data from the schools to conduct the random assignment, selecting students for the study sample. After random assignment, the research team will provide the list of students who

[1] We choose a randomization rate of one in three-and-a-half based on SAGA's ability to overrecruit qualified tutor applicants for open tutor positions.

were randomized to receive tutoring to SAGA staff and relevant school administrators at the schools and district. SAGA will offer program services to students randomized to treatment. Choosing not to participate in the program does not affect inclusion in the study sample, as researchers will examine administrative data for all students assigned to one of the groups in order to measure the "intent to treat." If it is determined that some schools did not comply with random assignment in a particular school year (based on the “first-stage” relationship between random assignment and actually receiving treatment), we will report results excluding these schools as a robustness check.

Lastly, to conduct student-tutor randomization, we will use existing administrative data from SAGA Innovations. SAGA will provide us with information on which student pairs will work together for the entirety of the academic year, as well as any restrictions for students that would affect which tutors they could work with (e.g. if students need a Spanish-speaking tutor, or if students need a tutor who can tutor an advanced mathematics class). SAGA will also provide us with information on tutors that are hired at each site (e.g. names, as well as tutor characteristics that could affect student-tutor pairings, such as whether a tutor speaks Spanish or can tutor advanced algebra). Student pairs will then be randomized to tutors, while taking any of the restrictions described before into account.

August 5th, 2020: In our preliminary analysis of the student-tutor randomization, we discovered that two schools did not randomly assign students to tutors in one school year each. The first-stage relationship between assigned tutor rank and actual tutor rank was negative, violating the relevance requirement for a valid instrument. Our main analysis of the relationship of student outcomes to tutor rank will exclude these two school sites from the analysis in the year they did not randomize students to tutors. We will report a version of the analysis including these sites as a supplementary analysis.

V. DATA SOURCES

The research team plans to use school administrative data from the Chicago Public Schools and from the New York City Department of Education to answer our research questions. Specifically, researchers will look at the effects of the program on math standardized test scores (our primary outcome of interest) that students take at the end of the year. These exams are administered to all students in the district.

The research team will also look at the impacts of the program on arrest rates in the subset of schools in Chicago where students are randomized to treatment. Our research team will utilize arrest data compiled by the Chicago Police Department for this analysis.

Lastly, the research team will also carry out observations in a sample of tutoring sessions during the school year to monitor implementation fidelity.

VI. OUTCOMES OF INTEREST

Below, our research team lists our primary and secondary outcomes of interest, and how key variables will be defined.

Primary Outcome Measures:

- Difference in math achievement
  - Performance on math standardized achievement tests, obtained from Chicago

    - Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year

Secondary Outcome Measures:

- Difference in math achievement

    - Performance on math standardized achievement tests, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: Two and three-years post intervention
- Difference in math course grades
    - Math course grades, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in non-math course grades
    - Non-math course grades, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in standardized test score achievement
    - Performance on additional sections of standardized tests (i.e. reading)
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in absentee rate
    - Number of school absences, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in index of schooling outcomes
    - Index of standardized (in Z-score form) outcomes for school persistence, absences, and course grades, obtained from Chicago Public Schools and New York City Department of Education administrative data
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in student misconduct
    - Number of school misconduct infractions, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in total courses failed
    - Number of total school courses failed, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in math courses failed
    - Number of math courses failed, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in non-math courses failed
    - Number of non-math courses failed, obtained from Chicago Public Schools and New York City Department of Education administrative database
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in school persistence
    - Measure from CPS and NYC DOE student records of school persistence (enrollment or graduation status by end of academic year)
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in violent crime arrests

    - Number of violent crime arrests, obtained from Chicago Police Department, Illinois State Police, New York City Police Department and New York State Police administrative databases (if available)
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in other arrests (property, drug, and other crimes)
    - Number of non-violent crime arrests, including property crimes, drug crimes, and other crimes, obtained from Chicago Police Department, Illinois State Police, New York City Police Department and New York State Police administrative databases (if available)
    - Time frame: End-of-year, as well as two and three-years post-intervention
- Difference in high school graduation rate
    - Difference in four-year and five-year high school graduation rates, obtained from Chicago Public Schools and New York City Department of Education administrative data
    - Time frame: Three, four, and five years after 9th and 10th grade enrollment
- Difference in college enrollment rate
    - Difference in college enrollment data, obtained from Chicago Public Schools and New York City Department of Education administrative data
    - Time frame: Three, four, five, six, seven, and eight years post-high school enrollment

VII. METHODS

To measure treatment effects for research question #1, the research team will estimate both intent to treat (ITT) and treatment on the treated (TOT) impacts. Researchers will estimate the ITT effect as follows:

$$Y=\beta_0+\beta_1 T+\beta_2 X+\varepsilon$$

where Y is the outcome of interest, T indicates students that are randomly assigned to be offered the chance to participate in the tutoring program, X is a set of baseline controls (which specifically includes randomization block, gender, age, learning disability, free/reduced lunch status, race, baseline grade level, GPA, number of As/Bs/Cs/Ds/Fs in the prior year, math and reading standardized test scores from the prior year, days absent from school, disciplinary incidents including suspensions and arrests, and a binary flag for students with missing GPA and attendance data), ε is a random error term, and $\beta 0$, $\beta 1$, $\beta 2$ are parameters to be estimated. The random assignment of T assures that under standard assumptions, Ordinary Least Squares (OLS) estimation yields an unbiased estimate of the ITT as the estimate of $\beta 1$, or the effect of being offered participation in the SAGA tutoring program. As all students who were randomized into the program were paired with tutors that were randomized via the process described above, our ITT (and subsequently, TOT) effect will specifically measure the effect of the program when administered at three-and-a-half times the scale as it is currently being administered.

The ITT measures the effect of being offered the chance to participate. As students assigned to treatment are not required to participate in the program, the ITT may not measure the effect of participation. The research team will measure the effect of participation using random assignment of T as an instrument for participation. If all participants were randomly selected (i.e. if there are no control students who are allowed to participate in the tutoring program) this method calculates the effect of the treatment on the treated (TOT), or the effect of participating for the group of students who choose to participate.

To gain insight into research question #2 above, which seeks to determine the relationship between program scale and effects, tutors at all sites are ranked by SAGA leadership based on relative expected quality. The research team will then randomize student pairs to tutors in an effort to identify a tutor's effect on student outcomes. Using this methodology, researchers can study whether tutor ranking predicts the size of the program effects. As the research team assumes that the program would hire tutors in the order of their ranking depending on the number of tutor slots they needed to fill, this analysis will shed light on the relationship between scale and effectiveness.

To analyze the impact of being assigned to a tutor of a particular ranking (i.e. the ITT estimate of research question #2), investigators will run regression models that regress academic outcomes on tutor rank. Our main outcome of interest is math standardized test scores.[1] Our primary analysis will model outcomes as a linear function of tutor rank. As a secondary exploratory analysis, we will estimate the shape of the relationship between outcomes and tutor rank using the following leave-one-out cross-validation exercise:

1. Estimate the relationship non-parametrically by running a regression at the student-level of the outcome (primary: math standardized test scores) on a full set of separate fixed effects for every tutor rank. Call the coefficient on the fixed effect of the r^th ranked tutor, $\hat{\gamma}_r$.

2. For rank r=1,…,R: Estimate using the student-level data the relationship between the outcome and tutor rank as a polynomial of order p=0, 1, 2, …, 10 holding the students assigned to tutors ranked r out of the sample. Use the coefficients from the polynomial terms to predict $\hat{\gamma}_r$, and call that $\hat{f}_p(r)$. Save the squared error for each polynomial: $\left(\hat{f}_p(r) - \hat{\gamma}_r\right)^2$.

3. Select the polynomial p that minimizes $\sum_{r=1}^{R}\left(\hat{f}_p(r) - \hat{\gamma}_r\right)^2$.

4. Report the estimated relationship between the outcome and rank using the selected order of polynomial.

In addition to the functions of tutor rank, each regression will include block fixed effects that capture how student pairs were randomly assigned to tutors. The blocks include student groups within a classroom with shared special restrictions (e.g. having no restrictions, needing a Spanish-speaking tutor, or needing a tutor who is qualified for advanced mathematics courses).[2] Other covariates in the model will be the same as those included in our ITT and TOT analysis for research question #1, noted above, to measure the effect of the program when administered at three-and-a-half times the scale as it is currently being administered.

As students switch tutors for a variety of reasons, researchers will also need to calculate the

[1] We will look at other academic outcomes – such as math course grades and GPA – as secondary outcomes of interest.

[2] Because the randomization was of student-pairs to tutors, randomization blocks can be defined either by student groups that were assigned with an identical probability distribution over tutors (as we do in our main specification) or by tutor groups that were assigned with an identical probability distribution over students. Researchers will report results controlling for block fixed effects using this alternative definition as a robust check.

TOT estimate to look at the impact of being assigned to and actually working with a tutor of a particular ranking. To do so, investigators will use the rank of the randomly assigned tutor as an instrument for the weighted average tutor rank, where the weight placed on each tutor's rank is equal to the proportion of time (measured in days) the student spent with that tutor. Investigators will use daily attendance data to create this weighted average. This method will help investigators understand whether working with a higher-ranked tutor means that a student actually receives better quality instruction, or whether an additional relationship between tutor rank and actual quality exists.

We will use only observed outcomes for the analysis. If an outcome is missing for more than 5 percent of the sample, we will also report the treatment effect on whether or not the outcome is observed. When baseline covariates are missing, we will impute missing values with zero and include an indicator for missingness as an additional baseline control.

VIII. CORRESPONDENCE WITH ETHICAL STANDARDS FOR RESEARCH

The research protocol has been approved by the University of Chicago Social & Behavioral Sciences Institutional Review Board, and will be reviewed by the Chicago Public Schools for approval prior to research activities taking place. Privacy of all data will be maintained through the University of Chicago Education Lab's extensive security procedures. The University of Chicago will obtain permission to use existing administrative data from the Chicago Public Schools and New York City Department of Education.