Learning without Teachers? Evidence from a Randomized Experiment of a Mobile Phone-Based Adult Education Program in Los Angeles

Christopher Ksoll, Jenny C. Aker, Danielle Miller, Karla Perez, Susan L. Smalley*

July 2014

Abstract: Over 755 million adults worldwide are illiterate. The spread of information and communication technology offers new opportunities to serve the educational needs of these populations. Using data from a randomized experiment of a mobile- phone-based adult education program (*Cell-Ed*) in Los Angeles, we find that students' reading scores are substantially increased over a four-month period, equivalent to a 2-4 year increase in reading levels. These results are robust to correcting for non-random attrition using a variety of non-parametric methods, including using the phase-in design to tighten the Lee (2009) bounds. The program also increased participants' self-esteem by 7 percent.

JEL codes: D1, I2, O1, O3

* Christopher Ksoll, School of International Development and Global Studies, University of Ottawa, 120 University, Ottawa, ON, Canada; christopher.ksoll@uottawa.ca. Jenny C. Aker, Department of Economics and The Fletcher School, Tufts University, 160 Packard Avenue, Medford, MA 02155; Jenny.Aker@tufts.edu. Danielle Miller, Dornsife College of Letters, Arts and Sciences, University of Southern California and Cell-Ed. Karla C. Perez-Mendoza, Graduate School of Education and Information Studies (GSE&IS), University of California Los Angeles; kcperez77@ucla.edu. Susan L. Smalley, Department of Psychiatry and Behavioral Science, UCLA and Cell-Ed; sue@suesmalley.com. We thank Michael Clemens, Jose Galdo, Louis-Philippe Morin, Concepcion Valadez, Bruce Wydick, as well as seminar participants at the University of San Francisco and Tufts University for helpful comments. We thank Ariadna Cornei, Stephanie Del Mel and Wei Yang for excellent research assistance. Ksoll acknowledges funding from a Canadian Social Sciences and Humanities Research Council (SSHRC) Insight Grant.

Education is positively correlated with many dimensions of well-being, both at the individual and aggregate levels. Yet educational achievements are still remarkably low in many parts of the world: It is estimated that there are over 755 million adults worldwide who are unable to read and write in any language (UNESCO 2012). This is not merely an issue that is confined to developing countries: In the US, approximately 14 percent of adults are considered to be functionally illiterate (National Center for Education Statistics 2003).[1]

Adult education programs have the potential to bridge this gap, but they are often characterized by low enrollment and high drop-out rates, potentially because the opportunity costs associated with such programs are too high for adults with families and busy work schedules. Moreover, these opportunity costs might not yield sustainable returns, as evidence suggests that adult learners' skills often depreciate after completing the program. The failure for literacy gains to persist may be due to the irrelevancy of such skills in daily life or limited opportunities to practice such skills in the learner's native language (Aker, Ksoll and Lybbert 2012).

More than four billion individuals have access to mobile phones, with developing countries representing some of the fastest-growing markets (ITU 2013). The widespread growth of mobile phone coverage has the potential to facilitate skills acquisition of illiterate adults, as well as increase the scope and scale of such programs. While a number of technology-based adult education programs have been developed, most require Internet access or smart phones, which are often not easily available to poorer populations in both developed and developing countries. Furthermore, educational programs delivered

[1] A person is considered illiterate if they are unable to read and understand basic prose in any language) A person is considered functionally literate if they can "engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing and calculation for his own and the community's development." (UNESCO 2005)

via information technology are often complements to teachers, rather than substitutes. As such, they are heavily dependent upon teacher availability and quality, a major constraint in many contexts.

We report the results of a randomized adult education program in Los Angeles, where a mobile phone-based adult education program in Spanish (*Cell-Ed*) was offered to illiterate adults. Unlike many other technology-enhanced education programs, the Cell-Ed learning curriculum was completely provided via a series of voice and SMS-based operations on the mobile phone, and therefore did not require teacher instruction or in-situ learning. Seventy Spanish-speaking adult students were randomly assigned to the treatment (*Cell-Ed*) or control group, with the control group phased into the program after a three-month period.[2] We find that the program substantially improved learning outcomes: The increase in *Cell-Ed* participants' reading skills after four months was equivalent to the reading skills children acquire after 2–4 years of schooling. We posit that these increases may be due in part to the flexibility of the curriculum, as learners opted to learn at all times of the day and for short durations, in stark contrast to the fixed schedules of many adult education programs.

A key claim of many adult education programs is that they empower students, primarily by providing them with the knowledge and skills necessary to acquire new labor market opportunities and live more independent lives. We investigate whether the *Cell-Ed* program affected empowerment, as measured by the Rosenberg self-esteem score (RSES) and the General Self-efficacy Scale (GSES). Our results provide more speculative evidence that the *Cell-Ed* program increased participants' self-esteem scores. Using real-time data on learning progress on the platform and weekly measures of self-esteem, we also provide insights into the dynamic relationship between learning progress (or lack thereof) and self-esteem.

[2] The minimum time spanned by different episodes of access to the platform is three weeks.

A potential threat to the internal validity of our findings is differential attrition between the treatment and control groups. To deal with non-random attrition, we implement two non-parametric approaches (Manski 1990, Lee 2009). Our first approach employs Lee's methodology, exploiting the randomized phase-in design of the program to further tighten the traditional Lee bounds by identifying the control group respondents and non-respondents after they have been phased into the program.[3] As a second approach, we combine the monotone treatment response (MTR) and monotone treatment selection (MTS) assumptions to bound the treatment effect (Manski 1997, Manski and Pepper 2000). Using these corrections, we are able to conclude that the *Cell-Ed* program is highly effective in improving reading levels, even when imposing minimal assumptions. Our paper thus proposes an alternative set of assumptions for bounding the treatment effect and suggests that phase-in designs may confer important identification benefits.

Our results contribute to the growing debate on the effectiveness of ICT-assisted learning in other contexts. While Barrera-Osario and Linden (2009) and Fairlie and Robinson (2013) find that computers have either no or mixed effects on learning outcomes, Banerjee et al (2007) found that computers increased students' math scores and were equally effective for all students. Barrow, Markman and Rouse (2009) find that students randomly assigned to a computer-assisted program obtained significantly higher math scores, primarily due to more individualized instruction. Most relevant to our study, Aker, Ksoll, and Lybbert (2012) find that learning how to use mobile phones in adult education classes in Niger increased test scores by .19-.25 s.d. Our paper provides additional evidence that a simple, self-directed educational device can lead to substantial improvements in learning, on the order of magnitude of 1.3-3.2. s.d.

[3] DiNardo et al (2006) show an alternative approach to tightening Lee bounds. This entails creating two groups with different survey attrition rates by randomizing the levels of effort exerted in recontacting survey respondents.

Our paper also speaks to the potential relationship between education and empowerment. The UNFPA, for example, states that "Education is one of the most important means of empowering women with the knowledge, skills and self-confidence necessary to participate fully in the development process", a sentiment that has often been repeated by donors, policymakers and practitioners alike. Yet what is surprising about these claims is the relative paucity of evidence to support them. There is significant literature measuring the impact of education programs on educational and labor markets outcomes, as well as health and fertility (see, for example, Case 2005). A separate strand of literature assesses the relationship between empowerment and economic development (see, for example, Duflo 2012, Doepke and Tertilt 2014) or the impact of different interventions on intra-household decision-making and empowerment (Ashraf 2009, Ashraf et al 2010). Yet there are few studies that explicitly link education and empowerment. Our paper contributes to this literature by measuring the impact of an educational program on self-esteem, as well as the dynamic relationship between learning and self-esteem over time.

The remainder of the paper is organized as follows. Section I provides background on the setting and the research design. Section II describes the different datasets and Section III outlines the estimation strategy. Section IV discusses the results, whereas Section V looks at the potential learning mechanisms. Section VI concludes.

I. Research Setting and Design

While the United States is one of the wealthiest countries in the world, it is estimated that 1 in 7 U.S. adults are functionally illiterate, defined as being unable to read and understand basic prose (National Center for Education Statistics 2003). The Hispanic population accounts for a disproportionate share of the functionally illiterate, with approximately 44 percent of Hispanics having literacy levels considered to be at or "below basic" level. Hispanic populations also represent 63% of the U.S' "least literate" adults, defined as those who fail the simplest screener questions.[4]

Greater Los Angeles has the highest rate of so-called "undereducated adults" in any major U.S. metropolitan area (Literacy Network of Greater LA and United Way 2004).[5] Similar to national figures, a disproportionate number of the low-literate and illiterate populations are immigrants, primarily Spanish-speaking immigrants who are unable to read or write in Spanish or English.

In an effort to address this issue, the U.S. Adult Education and Family Literacy Act provides approximately US$ 600 million per year in funding for states to implement adult education programs. A majority of these adult education programs are taught using teacher-based personal instruction, often in "adult schools" or community centers. Despite these efforts, adult education programs reach fewer than 16 percent of the low-literacy population in Los Angeles, and drop-out rates are well over 50 percent (Literacy Network of Greater LA and United Way 2004).

A. Cell-Ed Intervention

The widespread growth of mobile phone coverage has the potential to facilitate skills acquisition of illiterate adults, as well as increase the scope and scale of such programs. While a number of technology-based adult education programs have been developed, most require Internet access or smart

[4] The National Assessment of Adult Literacy is a nationwide representative survey on literacy in the US. Survey respondents representing approximately 4 million people (around 1.5 percent) were unable to communicate in English or Spanish, so that the assessments could not be administered. The assessment first asks respondents 7 very simple literacy questions. If respondents fail this initial screening (about 3% representing about seven million people did), they participate in a survey for the "least literate adults". 63% of the respondents in this category are Hispanics. Of the three percent, only 57% could read letters, 46% could read words.

[5] Literacy Network of Greater LA and United Way (2004) define literacy and numeracy levels slightly differently from the 2003 National Assessment of Adult Literacy (National Center for Education Statistics 2003). In their definition, individuals with a literacy "Level 1" are unable to engage in basic daily tasks, such as find an intersection on a map. Approximately 32 percent of the population in Greater LA (or 2.3 million people) fall into this category, as compared with 20 percent at the national level.

phones, which are often not available among poorer populations in Los Angeles. *Cell-Ed* is a platform that provides basic educational instruction via simple mobile phones. The platform uses voice (audio) and SMS messages to deliver 437 adult education lessons (called "micro-modules") to learners. Each micro-module consists of three components: 1) audio instruction: an audio lesson on a particular concept (vowels, consonants, words), and varying from 1-3 minutes in length, is introduced when the learner calls a designated number; 2) written instruction: a SMS message reinforcing the voice lesson is sent to the participant; and 3) interactive quiz: a SMS question is sent to the participant asking them about the lesson that they recently learned, and the participant must text a response. A correct response to the question triggers the beginning of the next micro-module, whereas an incorrect response leads to a repetition of the same micro-module until the user succeeds. To activate the program and each micro-module, participants call the *Cell-Ed* phone number from their own mobile phone.[6] Students could access the program 24 hours a day, seven days a week, allowing them to learn when, where and how they wished.

The *Cell-Ed* curriculum for this study was based upon a traditional Spanish adult education program (LEAMOS!) developed by the Centro Latino in Los Angeles. The traditional LEAMOS! curriculum is comprised of 43 lessons and teaches students simple Spanish letter and word recognition, as well as reading and writing skills. The curriculum is typically taught by an in-classroom teacher and takes approximately 150 hours to complete over a four-month period, or 9 hours per week. The 43 LEAMOS! lessons were adapted into 437 micro-modules for the *Cell-Ed* platform, with each micro-module including recorded audio instructions, SMS messages and queries for interactive testing. In

[6] If Cell-Ed participants did not own a mobile phone, they were provided with a simple phone as part of the program. All participants (regardless of mobile phone ownership) were also provided with an unlimited voice and SMS plan if they did not already have one.

addition to the micro-modules, pre-recorded audio messages were sent to each learner to offer encouragement at various points in the process.

B. Experimental Design

Prior to the introduction of the program, *Cell-Ed* partnered with two schools and five community resource centers with large Hispanic populations in the Los Angeles area. A variety of recruitment methods were used to recruit potential participants, including informational flyers, door-to-door visits by parent volunteers and presentations at community meetings, school fairs and school events. Using these recruitment methods, we identified 250 individuals and conducted an initial screening process via phone. The screening process collected information on individuals' socio-demographic characteristics, reading ability, mobile phone access and eyesight. Participants were excluded from further participation if they were over the age of 80 years, could read sentences, or required eyeglasses and did not own them, thereby reducing the sample size to 124 participants.[7]

The initial 124 participants were scheduled for an additional in-person baseline evaluation, of which only 89 attended (71 percent). Those who accepted to participate in the baseline evaluation were provided an ID number and asked to complete a battery of assessments, including the Woodcock-Muñoz Achievement Battery.[8] Participants who scored sufficiently low on the reading test (with a score lower than "basic" in two of three reading sub-tests) and did not suffer from neurological disorders were considered eligible for participation in the study. These exclusions further reduced the eligible sample to

[7] This initial screening process implies that our reduced sample was less literate than the initial sample of 250 individuals, but had access to eyeglasses, if necessary.

[8] For each survey round, participants were provided with compensation in the form of a US$50 gift card to a local supermarket chain. In addition, there was a US$100 incentive for completing the program in 4.5 months Thus, a treatment participant that completed all survey rounds and finished the program in 4.5 months would receive US$ 200, whereas a control participant who completed all survey rounds and finish the program in 4.5 months (after being phased in) would receive US$ 250.

70 participants. After finishing the screening process, participants were informed of their treatment status via a previously sealed letter (linked to their initial ID) that indicated whether they had been assigned to the treatment or control condition. This letter was read to them.[9] The ID numbers had been divided into strata and randomly assigned to either the treatment (the *Cell-Ed* program) or the control group (to be phased into the *Cell-Ed* program approximately 3-4 months later) with equal probability.[10] The random assignment took place before the evaluation, and enumerators were not informed of participants' treatment or control status.

The treatment thereby followed a randomized phase-in design. After the initial baseline evaluation, those in the treatment group were asked to return for a second evaluation approximately 4.5 months later, in order to insure that they had sufficient time to complete the *Cell-Ed* program.[11] The control group was eligible to start the *Cell-Ed* program approximately three months after the baseline, after having completed a midline survey. A timeline of the implementation and data collection activities is provided in Figure 1.

Overall, the screening process outlined above suggests that our sample is less educated than the broader population of Spanish-speaking immigrants in Los Angeles, as they did not perform well on initial reading tests. In light of the low levels of literacy among our sample, participants reported that

[9] As the control group was informed that they would receive the treatment at a later time, this could have changed their behavior between the first and second survey rounds (the John Henry effect). If the control group exerted less effort in learning than they would have otherwise, our results could potentially provide an upper bound for the treatment effect. If, however, this letter encouraged the control group participants to exert greater effort in learning during the first round, then our results would provide a lower bound on the treatment effect. As reading skills are difficult to acquire without some type of external support, we do not think that the John Henry effect is a primary concern in this context.

[10] In order to ensure that balance between treatment and controls were achieved for earlier and later potential participants, we stratified the random assignment by pre-screening ID, in the following strata: 1-22,23-46,47-70,70-100,100+. Within strata the ex ante probability of selection into treatment was 50 percent. We control for strata fixed effects in all OLS as the ex post probability differed due to ineligible applicants.

[11] Early piloting of the program suggested that most participants could complete the program in a 4.5 month time period. This time period was also similar to the time frame for most Leamos! learners.

they relied on families and friends to complete tasks that required reading and writing or simply avoided such tasks altogether. Most participants were employed in the informal economy, and stated that they were employed in jobs in which literacy was not essential.

II. Data

This paper relies upon four primary datasets. First, we administered comprehensive reading tests to measure the impact of the *Cell-Ed* on learning outcomes. Second, we collected data on student and household characteristics, as well as some qualitative data on their experiences as in Los Angeles. Third, we conducted short weekly phone calls to both treatment and control participants. Fourth, we collected real-time *Cell-Ed* usage data, providing user-specific information on how much time was spent on the program, as well as performance statistics (i.e., whether the student failed or passed a specific module).

A. Test Score Data

Reading tests were administered to all participants during the baseline, providing a baseline sample of over 89 students (of which 70 were retained for the survey, as outlined above). We administered a second round of reading tests for both groups approximately 3-4.5 months after the baseline, with a third round of follow-up tests for the control group only approximately 4.5 months after they started the *Cell-Ed* program. The second set of test scores allows us to measure the impact of the *Cell-Ed* program on learning, as the control group had not yet started the program. The third set of test scores for the control group allows us to assess whether the impacts were similar in the control group as they were in the initial treatment group.[12]

[12] For treatment individuals who completed the program in less than 4.5 months, the first follow-up survey was done as soon as possible after completion of the program. The main purpose of the control group was to ascertain how outcomes change over time in the absence of treatment. Typically, follow-up surveys would be implemented at the same time for both the

We administered the Spanish language equivalent of the widely used Woodcock-Johnson battery of literacy tests, the *Woodcock-Muñoz III Language Survey (WMLS-III).* The typical survey includes a screener (three tests) and a more-comprehensive seven-test battery, although our survey did not include the former.[13] We administered the cognitive battery of tests, which provides a composite verbal IQ score, as well as the reading achievement battery of tests, which were used to calculate two composite scores: the *basic reading score*, which is comprised of two sub-tests (letter-word identification and word attack); and the *broad reading score*, which is comprised of three sub-tests (letter-word identification, reading fluency and passage comprehension). The basic reading score covers literacy skills as narrowly defined, namely reading decoding and phonetic coding (identifying letters and building them up to words) (McGrew et al. 2007). The broad reading score can be interpreted as a measure of more advanced reading skills, as it also contains dimensions such as fluency and comprehension. Each composite score is calculated as an "age equivalent", which indicates the typical age of persons in the population who obtain a given score. For example, if a student's performance on the test of reading comprehension is equal to an age equivalent of 8.5 years, this means that his or her obtained *raw score* is equivalent to the predicted average for 8-year, 6-month old children in the norm group. We use the age equivalent throughout our paper since it has an intuitive interpretation.[14] As Spanish is a phonetic language, where each letter maps into one sound, becoming literate may be simpler than in English, where the mapping between letters and sounds is not unique (Abadzi, 2013).

treatment and control groups. In our case, the change over the course of 3 months in reading ability in the absence of treatment is zero, implying that this would be the case 4.5 months later as well, unless the comparison group learned how to read otherwise. Testing the treatment group when they finished the program (with a maximum date of 4.5 months) alters the interpretation of the estimate: It is not the impact on learning after 4.5 months, but the impact of *Cell-Ed* at the individually defined end of the program, censored after 4.5 months. Some learners continued after the 4.5 months, and this learning is not captured in our tests.

[13] The typical seven tests involve picture vocabulary, verbal analogies, letter-word identification, dictation, understanding directions, story recall and passage comprehension. Testing times range from approx. 25 minutes (screener) to 55 minutes (complete battery).

[14] We also provide results with z-scores based on the norming population of the tests.

B. Socio-Demographic Characteristics and Measures of Empowerment

The second primary dataset includes information on student and household characteristics. Among all eligible participants, we conducted a baseline survey, as well as a follow-up survey 3-4.5 months later. For the control group only, we conducted a third survey 4.5 months after their midterm survey. Each survey collected detailed information on students' demographics (age, gender, birthplace), education, employment, household size, mobile phone ownership and usage. We use these data primarily for balance checks and additional controls.

While collecting data on socio-economic characteristics, we also included separate survey questions on two measures of psychological dimensions of wellbeing: self-esteem and self-efficacy, as measured by the Rosenberg self-esteem scale and the general self-efficacy score. The Rosenberg self-esteem scale (RSES) is a series of statements, designed to capture different aspects of self-esteem (Rosenberg 1965). Five of the statements are positively worded, while the other five statements are negatively-worded. Each answer is assigned a point value, with higher scores reflecting higher self-esteem (the maximum is 30). The General Self-Efficacy Scale (GSES) (Schwarzer and Jerusalem 1995) is also a 10-item psychometric scale that is designed to assess whether the respondent believes he or she is capable of performing new or difficult tasks and to deal with adversity in life. The scale ranges in value from 12-60, with higher scores reflecting higher perceived self-efficacy. We use these results to measure the impact of the *Cell-Ed* program on participants' perceptions of empowerment.

C. Weekly phone calls

In addition to the in-person surveys, we contacted all participants via weekly phone calls. Treatment participants were asked whether anyone assisted them with *Cell-Ed* that week, whether they had technical difficulties, what they liked and disliked about the curriculum and whether there were any impediments to their studies. They were also asked about mobile phone usage and self-esteem. As the phone call might be intervention in itself, phone calls were also made to the control group. Control participants were asked the same questions on mobile phone usage and self-esteem, but not about the *Cell-Ed* program.

D. Cell-Ed Usage Data

The final dataset is comprised of records from the *Cell-Ed* platform, which logs every interaction between the student and the platform. This log includes the date and time the participant called in, the mini-module they accessed, the timing and content of SMS messages received and sent and their performance on the test question for that micro-module. Using these data, we calculate a number of statistics, including the number of days between different modules; how the student performed on each module (including the number of attempts to obtain a correct response); how quickly each student completed the program; and whether the student stopped calling (and hence dropped out of) the *Cell-Ed* program. We use these data to provide insights into the mechanisms through which the *Cell-Ed* program affected learning outcomes.

E. Pre-Program Balance

Table 1 provides an overview of student characteristics by treatment status. The evidence suggests that the randomization was successful in creating comparable groups along observable dimensions. Differences in students' characteristics between the treatment and control group before the program started are small and not statistically significant (Table 1, Panel A). Participants were 47 years old, and a majority of participants were women. While 51% percent of participants had some schooling, the average duration of schooling was quite low: 1.22 years. 51 percent of households in the sample owned a mobile phone, broadly in line with other studies on mobile phone ownership among Hispanic populations with lower levels of education.[15] Among all of these variables, only the difference in mobile phone ownership is statistically significant at the 10 percent level, with a higher mean in the treatment group as compared with the control group. As mobile phone ownership could be positively correlated with the treatment and our outcomes of interest, we control for baseline levels of mobile phone ownership as a robustness check.

Panel B shows the means of key outcome variables (broad and basic reading age equivalent, the Rosenberg score and the self-efficacy score) by treatment status. During the baseline, participants had mean basic and broad reading scores of 6, suggesting that their reading levels were equivalent to a six year old. There was no statistically significant difference between the two groups. While the results of the Generalized Self-Efficacy score were similar across treatment and control groups, participants in the treatment group had slightly higher baseline levels of self-esteem, with a statistically significant difference at the 10 percent level. As this measure is potentially correlated with treatment and the outcomes of interest, we control for baseline levels of self-esteem in the robustness checks as well.

A key concern with the balance tests is that the absence of any statistically significant differences may be due to our limited number of observations, hence increasing the size of our standard errors. While this might be a potential concern, we note that most of the differences in means (with the exception of years of schooling) represent less than 20 percent of the mean of the control group, suggesting that the magnitude of the differences is not large between the two groups for most of the baseline characteristics.

[15] While data on mobile phone ownership among Spanish-speaking immigrants in Los Angeles is not readily available, a survey by the Pew Trust estimated that mobile phone ownership among Hispanic households with lower income and educational levels ranged from 56-77%. http://www.pewhispanic.org/2013/03/07/closing-the-digital-divide-latinos-and-technology-adoption/.

F. Non-Response and Drop-outs

Table 2A shows the rates of survey non-response by treatment status and across all survey rounds. While definitions of survey non-response (attrition) and drop-out can differ, in our case, dropping out of the program at some point was correlated with eventual survey non-response, and so we treat them as interchangeable (and herein refer to them as non-respondents). Drop-outs are comprised of: 1) treated participants who could not make it past the first ten lessons; [16] 2) treated participants who made it past the first ten lessons but refused a follow-up interview; or 3) control participants who refused a follow-up interview, either because they refused a follow-up interview or dropped out after the first ten lessons (after being phased into the program).

Overall, survey non-response was 17 percent during the first round of the program. Yet there are significant differences by treatment status: Twenty-five percent of the treatment group did not participate in the follow-up survey as compared with 8.8 percent in the control group, with a statistically significant difference between the two. The rate of drop-out was higher during the second period (once the control group started the program), where average non-response was 32 percent. While these rates of attrition are high, we note that they are lower than what is typically observed in most adult education programs. The Literacy Network of Greater LA and the United Way (2004) estimate that 50 percent adult education students drop out within the first three weeks of such programs, with similar rates reported by Romain and Armstrong (1987). [17]

Table 2B presents more detailed information on learning and attrition. All students spent some time on the *Cell-Ed* platform, although there are marked differences in the duration of time spent by

[16] Follow-up data were not collected from this group of non-respondents, and hence that data cannot be included in the analysis.

[17] Literacy agencies covering almost 600,000 learners responded to their survey, of which only 55 percent tracked information on learner retention.

respondents (ie, those who completed the surveys) and those who eventually dropped out (non-respondents). Focusing on the treatment group (Panel A), respondents interacted with the platform over the course of 123 days (four months), as compared with 92 days (three months) for those in the treatment group who eventually dropped out (non-respondents). Respondents in the treatment group spent approximately 50 hours (3144 minutes) on the platform, with a minimum of 8 hours and a maximum of 148 hours. By contrast, treatment non-respondents spent only 8 hours on the platform, with a minimum of 2 hours and a maximum of 30 hours. Treatment respondents completed 402 modules (out of 437), about 90 percent of the curriculum, whereas treatment non-respondents completed only 31 micro-modules (7 percent of the curriculum). Once the control group phased into the *Cell-Ed* program, the difference in usage patterns between respondents and non–respondents follows a similar pattern (Panel B). Unsurprisingly, these results suggest that there are substantial differences in the *Cell-Ed* usage patterns among respondents and non-respondents, the latter of whom are primarily comprised of people who dropped out of the program.

The difference in means for the baseline characteristics of respondents and non-respondents are presented in the first three columns of Table A1, with the estimates from a logistic regression of non-response in Table A2.[18] Overall, non-respondents were largely similar along observable dimensions as compared with respondents, although part of this could be explained by our limited power to detect a statistically significant difference.[19] Non-respondents were more likely to have attended formal school and scored slightly higher on the baseline self-efficacy score (Table A1). These correlations are confirmed when looking at the determinants of drop-out: Attriters were more likely to have attended

[18] Columns 3-5 of Table A1 show the differences in baseline characteristics between treatment and control respondents during the initial phase of the program, before the control group was phased into the program.

[19] Looking at a broader range of covariates, initial treatment non-respondents differed at the 10% level from respondents on 8 variables out of 68. There were no statistically significant differences at the 5% level.

school and have a higher self-efficacy score than non-attriters (Table A2). In addition to these covariates, Table A2 also suggests that older participants are more likely to attrit and males were less likely to drop out. These comparisons confirm that non-differential attrition is a threat to the validity of our results, a topic that we address in more detail below.[20]

III. Estimation Strategy

All of the participants who were initially assigned to the *Cell-Ed* platform were provided access to the platform. The minimum time between the first and last contact with the platform among all treatment observations (non-respondents included) was 42 days.[21] As a result, there was no imperfect compliance: All of those *assigned* to the initial treatment participated in the platform (even if they dropped out at a later time), and none of the initial control group participants accessed the *Cell-Ed* platform during the initial phase. As a result, our treatment assignment variable is equivalent to our treatment participation variable, and our intention to treat (ITT) effect is equivalent to the average treatment effect on the treated (ATT).

Let $test_i$ be the composite basic or broad reading test score attained by student i after the program. $Cell\text{-}Ed_i$ is an indicator variable for whether individual i is assigned to the Cell-Ed intervention ($Cell\text{-}Ed$=1) or the control ($Cell\text{-}Ed$=0) in the first period. θ_R are fixed effects that indicate the randomization strata, which were based on the ID number. $\mathbf{X}_i'$ is a vector of student-level baseline covariates, primarily gender, age and a proxy for IQ. Ignoring attrition, we first estimate the following specification:

[20] An additional potential threat to the validity of our findings is spillovers. We cannot rule out the possibility that some treatment and control group participants knew each other before the program. In this case, however, our treatment effect estimates would represent a lower bound on the treatment effect in the absence of spillovers.

[21] When the control group was phased in to receive treatment, the minimum number of days between first and last interaction was 21 days.

(1) $$test_i = \beta_0 + \beta_1 CellEd_i + \mathbf{X}'_{i0} + \theta_R + \varepsilon_i$$

The coefficient of interest is β_1, which captures impact of the *Cell-Ed* program on learning outcomes. We also modify this specification to include the baseline outcome variable (a value-added specification), covariates that were statistically significant at baseline (mobile phone ownership and self-esteem), as well as a difference-in-differences (DD) estimation strategy.

While randomization would normally imply that β_1 has a causal interpretation, this is only the case if we assume that non-respondents would have progressed in their learning as much as those who continued with the program.[22] As noted in the methodological appendix, an unbiased estimate of the ATT can be estimated under the assumption that non-response is random – or observations are missing at random - which is not a reasonable assumption in our context. Nevertheless, we first proceed under this assumption before using various techniques to deal with non-random attrition.

IV. Results

Figure 2A depicts the mean reading scores for treatment and control respondents for the first round (prior to the phase in of the control group) before and after the program. Before the program, both groups had basic reading scores corresponding to an age equivalent of 6.5 years, without a statistically significant difference between the two (Table A3). This suggests that participants had a first to second grade reading level before the program. After the program, *Cell-Ed* participants had basic reading skills of 11.5 years, suggesting that they moved from a first grade to a sixth grade reading level, with a statistically significant difference at the 1 percent level. For the more comprehensive measure of

[22] An alternative assumption would be if attrition were equally high in treatment and control groups and we assume that treatment did not affect dropout for any survey participant. Then the estimated parameter would have the interpretation of the average treatment effect on the treated. In our survey there is a statistical difference in attrition, so this alternative identification assumption is untenable.

reading scores, broad reading scores, *Cell-Ed* participants moved from a first-grade to a fourth-grade broad reading level, whereas the comparison group showed no improvements in reading during this time. Thus, the *Cell-Ed* participants had broad reading scores that were 2-3 years higher than those in the control group, with a statistically significant difference at the 1 percent level. These results are similar for the control group once they phased into the program (Table A3), suggesting that the treatment effect is stable over time.[23] The distributional effects of the program are shown in Figures 2B-2C, suggesting that the treatment group's reading score distribution shifted sharply to the right.

A. Impact of Cell-Ed Program on Learning Outcomes

Table 3 presents the results of equation (1) for survey respondents. Controlling only for the *Cell-Ed* program and the randomization strata, the *Cell-Ed* program increased students' basic reading scores by 5.65 years, with a statistically significant effect at the 1 percent level (Column 1). These effects are robust to the inclusion of baseline covariates, namely age, gender and IQ (Column 2), a value-added specification (Column 3), a difference-in-differences specification (Column 4) and controlling for individual fixed effects (Column 5). The effects are also robust to the inclusion of the covariates that were different at baseline, as well as using normalized test scores as the dependent variable (Table A4). Overall, these results suggest that participating in the *Cell-Ed* program increased students' basic reading scores by 4.76 - 5.84 years within a four-month period, with a statistically significant effect at the 1 percent level.

[23] As there are only small baseline differences between the initial treatment and control group, and no significant changes in the control group's learning outcomes between the baseline and follow-up survey, any difference in impacts between these two groups would point to technical difficulties with the platform, spillover effects or learning on the part of the implementers. Given that the program is highly standardized through the platform, the absence of a difference is not too surprising.

The estimates are smaller in magnitude for the broad reading scores, which capture a broader set of reading ability including fluency and comprehension: the *Cell-Ed* program increased broad reading scores by 2.58 years when controlling only for the treatment variable and strata fixed effects, with a statistically significant effect at the 1 percent level (Column 6). These effects are similar when controlling for baseline covariates (Column 7), a value-added specification (Column 8), a DD specification (Column 9) and controlling for individual fixed effects (Column 10). The results are also robust to the inclusion of the baseline mobile ownership and self-esteem scores and using normalized test scores as the dependent variable (Table A4). The smaller impacts of the program on broad reading scores relative to basic reading scores is not surprising, as the *Cell-Ed* program focused on basic reading skills, rather than fluency and reading comprehension.[24]

Overall, the results in Table 3 suggest that *Cell-Ed* significantly increased participants' reading outcomes over the four-month period, with an increase that is comparable to a child's increase in reading levels over 2.5 to 5 years. Comparing our effects with those of other rigorous evaluations of adult education outcomes is difficult, as such evaluations are relatively rare and often use different measures of learning. Nevertheless, a large scale evaluation of an adult education program in India finds that the program increased learning outcomes by .06 – 0.15 s.d. (Banerji et al 2013), whereas Aker, Ksoll and Lybbert (2012) find that a mobile phone-enhanced adult education program increased writing and math test scores by .20-.25 s.d. as compared with a traditional adult education program. When using normalized test scores (Table A4), our effects suggest a treatment effect of 1.3-3.9 s.d. for broad and

[24] Table A3 presents evidence that the initial control group benefits as much from the treatment when they are phased-in. Column 4 tests for difference in outcomes between treated treatment group respondents and treated control group phase-in respondents and cannot reject that they are equal.

basic reading scores, respectively, which is considerably larger than the typical impacts of adult education programs.

B. Dealing with Attrition

The previous results ignore the differential attrition between the treatment and control groups, and hence provide biased estimates of the impact of the *Cell-Ed* program. We first address attrition by bounding the treatment effects under the assumption of monotonicity and implementing Lee bounds, and then use the phase-in design of the program to further tighten these bounds.

Standard Lee Bounds

The initial treatment group in our context was surveyed over two rounds and is divided into two groups: respondents (with a sample size (N) of 27) and non-respondents (N=9). The control group was surveyed over three rounds, and hence can be divided into three groups: The "Always non-respondents" (those who dropped out of the survey after the baseline, N=3); the phase-in non-respondents (those who participated in baseline and second round, but dropped out by the third round, N= 10); and the phase-in respondents (or those control group participants within who participated in all three survey rounds, with N=21).[25] Figure 3A shows the phase-in design of our specific program, indicating the proportion of baseline respondents who participated in each round of the survey.

Lee bounds assume monotonicity, in other words, that the likelihood of non-response is monotonically related to receiving the treatment (Lee 2009). This is equivalent to stating that receiving the treatment either makes *all* observations (weakly) *more* likely to respond to the survey, or *all*

[25] There is no ranking of outcomes associated with this graph, as observations in the phase-in non-respondents could well have expected outcomes that are higher than the lower bound of the phase-in respondents.

observations (weakly) *less* likely to respond.[26] This assumption rules out the possibility that the treatment may affect different sub-groups differently, such as increasing the likelihood of response for one specific sub-group while decreasing the likelihood for another. This also implies that the non-respondents in the group with lower non-response would not have responded if their treatment status were changed. When attrition is higher in the treatment group, as is the case in our experimental set-up, the monotonicity assumption implies that the control observations who did not respond in the second survey would also not have responded had they immediately received treatment (Figure 3B).[27],[28]

Under these assumptions, we first estimate the upper and lower bounds by assuming the "best" and "worst" case scenarios.[29] In our case, the treatment group has an attrition rate of 25 percent and the control group has an attrition rate of 8.8 percent, so the difference in drop-out is 16.2 percentage points. To estimate the *lower bound*, we trim the lowest 17.7 percent of the *observations* from the control group. [30] The lower bound is thus equivalent to assuming that the outcomes of additional treatment non-respondents correspond to those control group observations who make up the bottom end of the control group distribution: We compare the observed observations in the treatment group (who make up 75% of all the treatment group observations) to the 82.3% "best" observed observations in the control group (who make up 75% of all control group observations) (Figure 3C). As is usual, the confidence intervals around this estimated *lower bound* are presented. We present bootstrapped confidence intervals around all bounds. To create the *upper bound*, we trim the highest 17.7 percent of observations from the control

[26] We write "weakly" because it is possible that the treatment leaves some observations' likelihood unchanged.

[27] To be exact, the term "always non-respondent" thus already embodies an assumption we are making.

[28] In the appendix, we discuss a weaker assumption than this.

[29]This assumes that the additional percentiles of non-respondents in the treatment group correspond to, respectively, the best and worst respective percentiles of observed outcomes in the control group (Lee 2009).

[30] Note that we calculate the proportion of observed control group observations that is trimmed as the difference in attrition divided by the number of control group observations that are observedis 17.7=16.2/(100-8.8). Because of indivisibilities, the exact trimming proportion is 19.35 in the empirical application.

group (equivalent to 16.2 % unconditionally). The upper bound is equivalent to assuming that the additional non-respondents are the top learners in the control group, and is constructed by comparing the observed treatment group outcomes to the 82.3% worst control group observations (who make up 75% of all control group observations) (Figure 3D).

Adjusted Lee Bounds

The phase-in design of the program allows us to observe which control group participants later dropped out of the program once they received the treatment (Figure 3A). Using this information, we further tighten the traditional Lee bounds by assuming that the treatment non-respondents would not have participated in the third round in the survey had they been assigned to the initial control group (Figure 3E), either because they are *always non-respondents* or because they would also not have responded after receiving treatment (Figure 3E). This rules out a case in which the treatment affects the probability of non-response in opposite directions for those in the treatment group and those in the control group when they become treated.[31],[32]

Using this additional assumption, we can then construct the lower bound on the treatment effect by discarding the worst cases among the phase-in non-respondents and comparing the mean outcomes of the treatment group with the mean outcomes of the newly-constructed control group (Figure 3F). The upper bound similarly discards the best cases among the phase-in non-respondents (Figure 3G). Compared with standard Lee bounds, these bounds are tighter, as we are trimming the same proportion

[31] As such, it is really implied by the monotonicity assumption, but we have to also assume that the monotonicity assumption holds across time. It can't be that treatment makes a control person more likely to respond just because he or she waited for three months for the program, rather than less as for anyone else.

[32] This is equivalent to assuming that the control group phase-in non-respondents are ordered in their likelihood of non-response between the "always non-respondents" (who respond in one survey) and the "phase-in respondents" (who respond in three surveys), *even if they had been in the treatment group*.

of respondents but in a smaller group. The methodological appendix contains the formal proof that the bounds from this methodology, herein called the "adjusted Lee bounds", are tighter than the standard Lee bounds.[33,34]

A phase-in design also has an additional advantage relative to standard experimental designs in that it most likely reduces the attrition in the control group if they are aware that they will receive the treatment later. As many social experiments are not blinded (i.e. particpants *do* know if they are receiving treatment or not), researchers often worry that attrition in the control group tends to be higher than in treatment groups, as treatment respondents may (incorrectly) associate the survey with the benefits of the program. By providing an incentive for control participants to remain in contact with the program, attrition may be reduced.

Bounding under the Monotone Treatment Selection (MTS) and Monotone Treatment Response (MTR) Assumptions

Using the phase-in design, we can also implement two alternative assumptions on the potential outcomes of non-respondents: The monotone treatment selection (MTS) assumption and monotone treatment response (MTR) assumption. The *monotone treatment selection* assumption (Manski and Pepper 2000) implies that program drop-outs have lower outcomes on average than those who take up the treatment. More specifically, it assumes that those who select into the treatment have higher

[33] Lee (2009) shows that conditioning on any discrete variable will weakly tighten the bounds. In our case, we prove it strongly tightens the bounds. To provide an intuiton why this leads to a tighter upper bound (Figure 3G), note that the mean grade of the top six students in a sub-group (control group phase-in non-respondents) is always lower than that of the top six in the larger group (all control group respondents), unless the top six students in the sub-group are also the top six students in the larger group. We can reject the latter exception with our data. Since we are subtracting the non-trimmed (who now have a higher mean) from the treatment group, we get a lower upper bound.

[34] When there is treatment effect heterogeneity, the bounded estimate measures the treatment effect on the treatment respondents. In our context, this is highly correlated with continuing the *Cell-Ed* program until the survey (and perhaps completing it) and being surveyed.

expected outcomes under both treatment and non-treatment than those who do not. This allows us to bound the outcomes (from above) of the control group drop-outs during the second round.[35] The *monotone treatment response* assumption (Manski, 1997), on the other hand, specifies that treatment cannot make anyone worse off. This can be implemented by bounding the missing treatment observations. Given the very high correlation between the baseline and follow-up outcomes in the control group (with a correlation coefficient of 0.9 and a small difference in means), this is a plausible strategy. We thus construct plausible lower bounds for the average treatment effect by implementing MTS combined with MTR, discussing the assumptions and methodology more thoroughly in the methodological appendix.[36]

C. Bounding Results

This section first provides the bounding results using the traditional and "adjusted" Lee bounds, and then provides lower bounds on the treatment effect using the MTS and MTR assumptions.

Table 4 shows both the traditional and adjusted Lee bounds for basic and broad reading scores. The relevant bounds on the treatment effects are in bold, namely, the lower bound of the 95% confidence interval (around the lower bound), as well as the upper bound of the 95% confidence interval (around the upper bound).[37] In all cases, the treatment effects are bounded away from zero: The point estimates of the lower bound for basic and broad reading scores are 4.43 and 2.08 years, respectively, with a statistically significant effect at the 1 percent level (Panel A, Columns 1 and 4). Focusing on the

[35] Specifically, we bound the second round outcome of the control group dropouts by including the larger of two values: the minimum of those who eventually took up treatment and the observations own baseline value. This is an even more conservative approach.

[36] The panel structure of the data allows for many different reasonable assumptions to be implemented. For example, we can implement MTR based on the changes due to the treatment; as well as bounding changes between two surveys in the control group dropout by changes of the observed changes in the control group among respondents who do not drop out.

[37] The lower bound of the 95% confidence interval of the lower bound and the upper bound of the 95% confidence interval of the upper bound are conservative estimates of the 95% confidence interval around the estimate (Lee, 2009). As noted above, we bootstrap the confidence intervals.

conservative lower bound of the 95% confidence interval, the treatment effect is bounded below by 1.58 and 0.72 years for basic and broad reading scores, respectively (Panel A, Columns 2 and 5).

As predicted, the adjusted Lee bounds are tighter (Panel B). The point estimates for the lower bound are 4.54 and 2.04 for basic and broad reading scores, respectively, with a statistically significant effect at the 1 percent level (Panel B, Columns 1 and 4). The lower bounds of the 95% confidence interval are 1.73 and 0.84 years (Panel B, Columns 2 and 5). Thus, making weak assumptions on the non-respondent outcomes, both the Lee and adjusted Lee bounds provide strong evidence of the positive and statistically significant impacts of the *Cell-Ed* program on students' learning.

Table 5 presents the lower bounds on the treatment effect using the MTS and MTR assumptions. The point estimates on the lower bound of the treatment effect are 3.31 and 1.54 years for basic and broad reading scores respectively, with a statistically significant effect at the 1 percent level (Columns 1 and 4). The lower bounds of the 95% confidence intervals confirm the patterns observed in Table 4, suggesting that *Cell-Ed* increased basic and broad reading scores by 1.07-.5 years, respectively. While considerably smaller in magnitude, these bounded estimates still suggest that the *Cell-Ed* program resulted in an 8-16% increase in baseline basic and broad reading scores over a four-month period.[38]

[38] The difference in the magnitude of the lower bounds in Tables 4 and 5 can be explained by two factors: the strength of the assumptions made on the drop-outs and the different estimations of the treatment effects. With regards to the latter, we note that the treatment group splits into two groups: 1) the treatment respondents, namely, the group who spent an average of 50 hours and completed at least 184 mini-modules. For this group, the *Cell-Ed* program "worked"; and 2) the group that spent on average 8 hours on the platform and completed a maximum of 104 mini-modules. For this group, the *Cell-Ed* "worked less". T he Lee bounds in Table 4 bound the treatment effect for the treatment respondents, and suggest that the point estimate for basic reading scores was 4.54, with a lower bound of 1.73. The MTS and MTR assumptions of Table 5 bound the average treatment effect for the whole population, and find a point estimate of 3.31 and a lower bound of 1.07 for basic reading scores.

D. Impacts of the Cell-Ed Program on Empowerment

The previous results have shown that the *Cell-Ed* program increased students' basic and broad reading scores in a relatively short period of time. Yet beyond learning, a key claim of many adult education programs is one of empowerment: educated adults are able to make better decisions, read street signs and prescriptions and search for jobs that require literacy skills. This section investigates whether the increased learning due to *Cell-Ed* led to improvements in self-esteem and self-efficacy, using the RSES and the GSES.

Table 6A reports the results of the same specifications as those in Table 3, using measures of self-esteem and self-efficacy as the dependent variables. Overall, the *Cell-Ed* program increased students' self-esteem score by 2.49 points, with a statistically significant effect at the 5 percent level (Table 6A, Column 1). These results are robust to the inclusion of additional covariates (Column 2), as well as a value-added specification (Column 3). However, the result is smaller in magnitude and no longer statistically significant once the DD (Column 4) and fixed effects specifications (Column 5) are used.

While overall self-efficacy scores are higher in the treatment group, there is no statistically significant impact of the *Cell-Ed* program on self-efficacy for the first three specifications (Columns 6-8). This is perhaps unsurprising, as self-efficacy scores were relatively higher in the control group at baseline. Using DD and fixed effects specifications (Columns 9 and 10), the *Cell-Ed* program is associated with an increase of 3.75-4.75 points in students' self-efficacy scores, with a statistically significant effect at the 10 percent and 1 percent levels, respectively.[39],[40]

[39] Table A5 shows that these results are robust to including baseline mobile ownership and self-esteem, as well as using normalized empowerment scores as the dependent variable.

As was the case with the test score results, the results in Table 6A are likely to be biased due to non-random attrition. Table 6B corrects for non-random attrition using the Lee and adjusted Lee bounds. Unlike the reading results, the results are not robust to using the Lee bounds: the lower Lee bounds for both empowerment measures are not statistically significant from zero. Using the adjusted Lee bounds, the lower bound for the basic reading score is statistically significant at the 10% level, while the confidence interval around the lower bound for broad reading scores includes zero.[41] Bounding the results using the MTR and MTS assumptions (Table 6C), the estimate of the lower bound of the *Cell-Ed* treatment effect on the self-esteem score is 2.3 and statistically significant at the 5% level, whereas the effect on the self-efficacy is not statistically significant from zero. Overall, these results suggest that the *Cell-Ed* program is associated with an increase participants' self-esteem, but we are unable to conclude that there is a statistically significant effect on self-efficacy.[42]

While Table 6 shows the impact of the *Cell-Ed* program on self-esteem and self-efficacy at the end of the program, we are also interested in the relationship between the *Cell-Ed* program, learning and self-esteem over time. Table 7 investigates this relationship using data from the weekly self-esteem measurements and the *Cell-Ed* platform. Overall, self-esteem is negatively associated with lack of learning progress during the previous week (Column 1). In other words, as a student's proportion of incorrect responses during the previous week increases, the student's self-esteem decreases. When disaggregating these results by gender (Columns 2-4), the relationship is driven entirely by men. This

[40] Table A6 uses an empowerment index comprised of the normalized RSES and the normalized GSES as the dependent variable. The results are broadly consistent with those in Table 6.

[41] The estimated lower bound for the index of empowerment is significantly different from zero at the 5% level of significance when implementing the adjusted Lee bounds, but is insignificant for the estimated Lee bounds.

[42] The estimated lower bound for the index of empowerment is significantly different from zero at the 5% level of significance when implementing the MTS and MTR assumptions.

suggests that while overall self-esteem is higher for *Cell-Ed* participants by the end of the program, perceptions of self-esteem may change over time, particularly when experiencing learning failures.

While the results in Table 7 provide important insights into the dynamic relationship between learning and self-esteem, Table A8 shows how learning is associated with the likelihood of dropping out of the program. *Cell-Ed* participants who had a higher proportion of errors over the course of their interactions with the platform were more likely to stop interacting with the platform and drop out of the subsequent survey round (Column 1).[43] Columns 2 and 3 focus on the sub-sample of women: Female participants with a proportion of errors exceeding a threshold of 40% were more likely to drop out of the program. The effects seem to be particularly strong among participants with a high proportion of errors, as suggested by the quadratic specification (Columns 4-6).[44]

These results, taken together, provide some additional insights into the relationship between learning, self-esteem and non-response in the context of our non-parametric bounds. First, the heterogeneous effects in Table 7 suggest that conditioning on gender would further tighten the bounds for the empowerment regressions in Tables 6B and 6C. Unfortunately, our sample size does not allow us to do this. Second, these results raise some questions about the validity of our bounding assumptions for the empowerment outcomes. For example, if a larger proportion of learning errors leads to lower self-esteem, and a learner is more likely to drop out after these errors have lowered his or her self-esteem below their baseline level, then the MTR assumption would not be valid. We do not believe that this is the case for men, as few men drop out and there seems to be no pattern linking men's errors to attrition (although we acknowledge the low power to detect such a pattern). Yet a dynamic in which

[43] The monetary incentive provided for completing the program may play a role.

[44] We note that the sample for men is too small for robust results, in particular, because only 12.5 percent of men (2 out of 16) drop out, as compared to almost one third for women.

women whose self-esteem was lowered by their lack of progress in the program and eventually drop out is consistent with the above findings. Table A9 investigates this further. We find that female *Cell-Ed* participants who eventually drop out do not have lower levels of self-esteem immediately prior to dropping out (as compared with their own baseline levels of self-esteem). This suggests that the threat to the validity of the MTR assumption is not a primary concern for our self-esteem bounding results.[45]

V. Mechanisms

Why might the *Cell-Ed* program improve students' learning outcomes, especially in the absence of a teacher? A key hypothesis of the program was that a mobile phone-based program might reduce the opportunity costs of investing in adult education for busy adults by allowing them to learn when, where and how they wished. We are able to test this hypothesis by using the *Cell-Ed* platform data, which provides information on when *Cell-Ed* students interacted with the platform.

Figure 4 shows the learning patterns of *Cell-Ed* participants. Overall, participants learned at all times during the day and night, at times much later than a standard adult education class would operate, suggesting that the mobile-based course was more appropriate for adults' work schedules (Figure 4A). Figure 4B suggests that *Cell-Ed* participants spent less time on the platform during weekends as compared with weekdays, perhaps due to their work schedules. Finally, Figure 4C shows the distribution of interactions with the platform in terms of the amount of time spent: most interactions are quite short, approximately 10 minutes, suggesting that learners use the platform for short learning episodes. This is in contrast to most adult education programs, which hold classes for several hours on a

[45] Regarding the reading outcomes, we posit that participants' reading skills could not actually decrease in response to an adult education treatment.

weekly basis. On average, though, time spent on learning is similar to standard literacy programs: Cell-Ed learners were spending an average of 2.95 hours per week (3.75 for respondents) on the platform. The target for in-situ learning is about 4-6 hours per week. Accounting for in-situ absences, the 50 percent dropout rate and travel time suggests that *Cell-Ed* learners spent slightly more time (on average) on literacy acquisition than students in a standard adult education program.[46]

While we do not know whether this flexibility is indeed the key to the large improvements in reading outcomes, the useage patterns highlight that people did make use of it.

VI. Conclusion

Information technology, and in particular mobile phones, enables individuals to access information at any time and location. Yet despite the potential, there is limited evidence that simple mobile phone devices can be used to teach basic educational skills, such as literacy and numeracy. Using a randomized control trials of a mobile phone-based adult education program (*Cell-Ed*), we find that the program significantly increased adult students' reading levels within a short period of time. These results are robust to correcting for significant non-random attrition using a variety of non-parametric methods. We also find that the *Cell-Ed* program is correlated with an increase in students' self-esteem by the end of the program, and that there is a dynamic relationship between learning progress and self-esteem.

On a methodological side, our paper contributes to the literature on the use of non-parametric approaches to bounding treatment effects. The phase-in design of the program allows us to observe

[46] Figure 5 shows learning patterns by baseline employment status. Employed participants have somewhat different patterns of accessing the platform, learning for longer periods on weekends and in the evenings as compared to their unemployed counterparts.

which control group participants later dropped out of the program once they received the treatment. It probably also contributed to reducing attrition in the control group. Exploiting the randomized phase-in design, we are able to tighten the Lee bounds. We are also able to show that this experimental set-up provides a number of identification possibilities that are absent with simple random assignment. An additional benefit of phase-in designs is that they address concerns by non-governmental partner organizations of excluding observations from treatment, which however, comes at the expense of longer-term treatment effects.

Admittedly, our experimental set-up has several limitations. First, we are unable to compare learning via the *Cell-Ed* platform with learning in a traditional adult education program, or the interaction between the two. As a result, we are unable to conclude whether such programs are complements or substitutes for teachers and in-classroom learning. Second, our small sample size greatly limits the external validity of our results. Nevertheless, our results show that a distance learning program via a simple mobile phone significantly improved adults' learning outcomes in this context, and suggests that there is great scale and scope for using these technologies in education programs in both developed and developing countries.

References

Abadzi, Helen. 1994. "What We Know About Acquisition of Adult Literacy: Is There Hope?," In *World Bank discussion papers,*, ix, 93 p. Washington, D.C.: The World Bank.

Abadzi, Helen. 2013. Literacy for All in 100 Days? A research-based strategy for fast progress in low-income countries," *GPE Working Paper Series on Learning No. 7*

Aker, Jenny C., Christopher Ksoll and Travis J. Lybbert. October 2012. "Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger." *American Economic Journal: Applied Economics.* Vol 4(4): 94-120.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics*, 3(3): 29–54.

Angrist, Joshua D., and Jorn - Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton, New Jersey: Princeton University Press, 2009).

Ashraf, Nava, Dean Karlan, and Wesley Yin. March 2010. "Female Empowerment: Further Evidence from a Commitment Savings Product in the Philippines." *World Development* 38, no. 3 (March 2010): 333–344.□

Ashraf, Nava. September 2009. "Spousal Control and Intra-Household Decision-Making: An Experimental Study in the Philippines." *American Economic Review* 99, no. 4 (September 2009): 1245–1277.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics*, 122(3), pp. 1235-64.

Rukmini Banerji, James Berry and Marc Shotland. 2013. "The Impact of Mother Literacy and Participation Programs on Child Learning: A Randomized Evaluation in India."

Barrow, Lisa, Lisa Markman and Cecilia Elena Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal: Economic Policy*, 1(1), pp. 52-74.

Blunch, Niels-Hugo and Claus C. Pörtner. 2011. "Literacy, Skills and Welfare: Effects of Participation in Adult Literacy Programs." *Economic Development and Cultural Change*. Vol. 60, No. 1 (October 2011): 17-66.

Carron, G. 1990. "The Functioning and Effects of the Kenya Literacy Program." *African Studies Review*, pp. 97-120.

Case, Anne. 2005. "The Primacy of Education," In *Understanding Poverty,* ed. Abhijit Vinayak Banerjee, Roland Benabou and Dilip Mookherjee. Oxford ; New York: Oxford University Press.

De Grip, A. and J. Van Loo. 2002. "The Economics of Skills Obsolescence: A Review," In *Research in Labor Economics,* ed. J. Van Loo and K. Mayhew, 1-25.

DiNardo, J., J. McCrary, and L. Sanbonmatsu. 2006. "Constructive Proposals for Dealing with Attrition: An Empirical Example." Working paper, University of Michigan**.**

Duflo, Esther. 2012. "Women Empowerment and Economic Development." *Journal of Economic Literature***.** 50(4). 1051-1079.

Doepke, Mathias and Michele Tertilt. 2014. "Does Female Empowerment Promote Economic Development?" *NBER Working Paper 19888,* NBER, Inc.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, 47 (1), 153–162.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–84.

Imbens, Guido and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters," *Econometrica*, November, 72 (6), 1845–1857.

Kim, Y., R. Telang, W. Vogt and R. Krishnan. 2009. "An Empirical Analysis of Mobile Voice Service and SMS: A Structural Model." *Management Science*, 56(2), pp. 234-52.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects" *The Review of Economic Studies*, 6, 1072-1102.

Linden, Leigh. 2008. "Complement or substitute? The effect of technology on student achievement in India." JPAL Working Paper.

Literacy Network of Greater LA and United Way. 2004. Literacy @ work. Skills Today, Jobs Tomorrow. Los Angeles: Literacy Network of Greater LA and United Way.

Manski, Charles. 1990. "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 1990, 80, 319–323.

Manski, Charles. 1997. "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

Manski, Charles, John Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling" *Econometrica*, 68, 997-1010.

McGrew, Kevin, Frederick Schrank and Richard Woodcock. 2007. *Techinical Manual. Woodcock-Johnson III Normative Update.* Rolling Meadows. IL: Riverside Publishing.

National Centre for Education Statistics. 2003. "The 2003 National Assessment of Adult Literacy", Washington DC, USA: National Centre for Education Statistics, U.S. Department of Education.

Ortega, Daniel and Francisco Rodríguez. 2008. "Freed from Illiteracy? A Closer Look at Venezuela's Mision Robinson Literacy Campaign." *Economic Development and Cultural Change*, 57, pp. 1-30.

Osorio, Felipe, and Leigh L. Linden. 2009. "The use and misuse of computers in education: evidence from a randomized experiment in Colombia." *The World Bank Policy Research Working Paper Series.*

Oxenham, John, Abdoul Hamid Diallo, Anne Ruhweza Katahoire, Anne Petkova-Mwangi and Oumar Sall. 2002. *Skills and Literacy Training for Better Livelihoods: A Review of Approaches and Experiences*. Washington D.C.: World Bank.

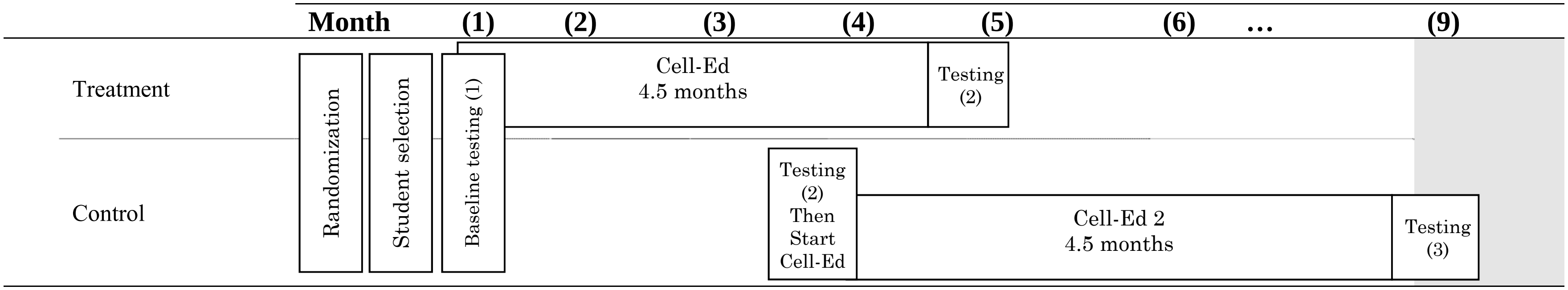
Romain, R. and L. Armstrong. 1987. *Review of World Bank Operations in Nonformal Education and Training*. World Bank, Education and Training Dept., Policy Division.

UNESCO. 2005. *Education for All: Global Monitoring Report. Literacy for Life*. Paris: UNESCO.

UNESCO. 2008. *International Literacy Statistics: A Review of Concepts, Methodology and Current Data*. Montreal: UNESCO Institute for Statistics.

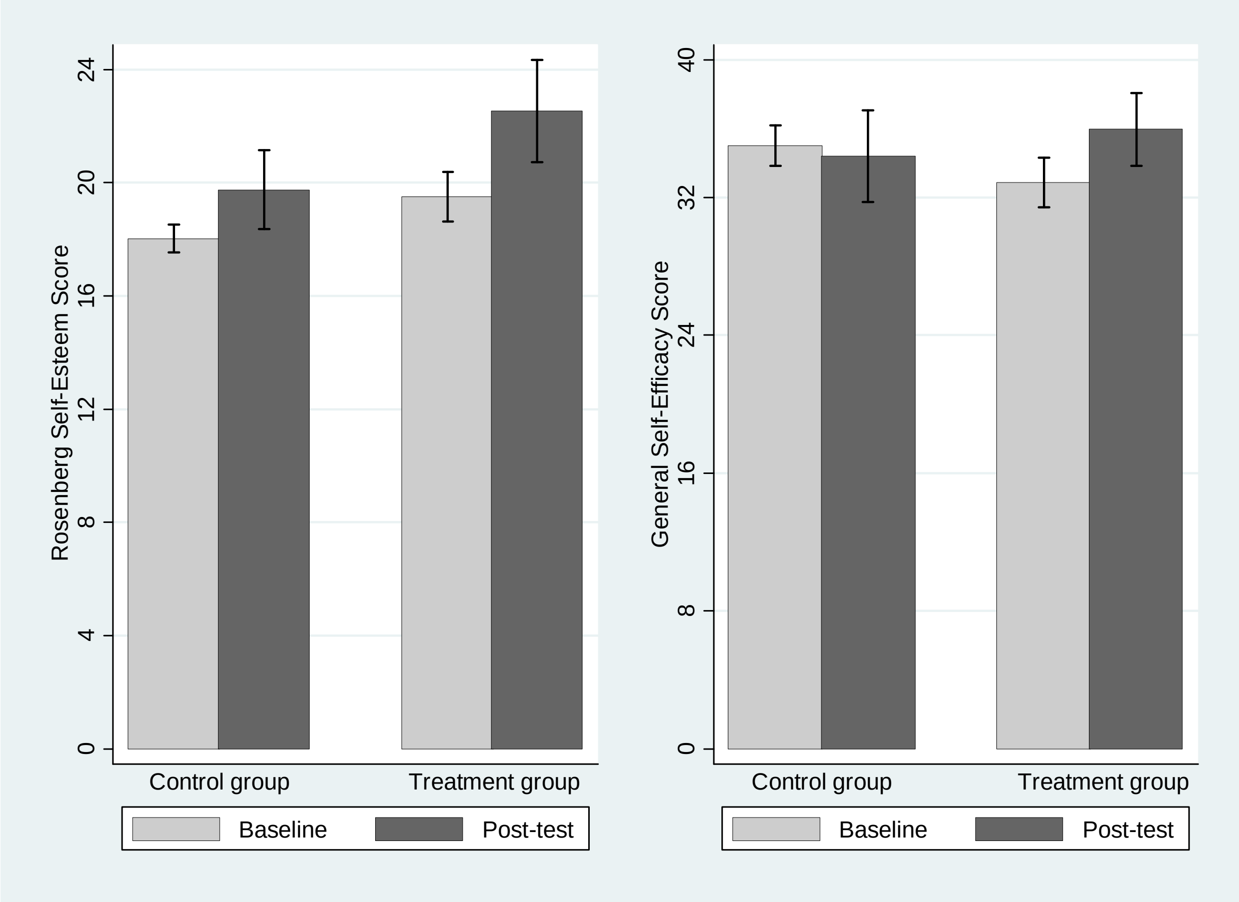
UNESCO. 2012. *Education for All: Global Monitoring Report. Youth and Skills: Putting Education to Work*. Paris: UNESCO.

Figure 1. Timeline of Data Collection and Adult Education Activities



Notes: This figure represents the timeline for the *Cell-Ed* program. Testing (1), Testing (2) and Testing (3) refer to the first, second an third round of testing, respectively.

Figure 2A: Basic and Broad Reading Scores (by Treatment Status and Survey Round)



Notes: Graph pictures the means scores on the Woodcock-Munoz III reading assessment, for basic (left panel) and broad (right pa reading scores, by treatment status. Baseline survey measurement means are in lighter color, Round 2 measurements in darker col depicted are the 95% confidence intervals around the mean.

Figure 2B: Distribution of Basic Reading Score by round and treatment status

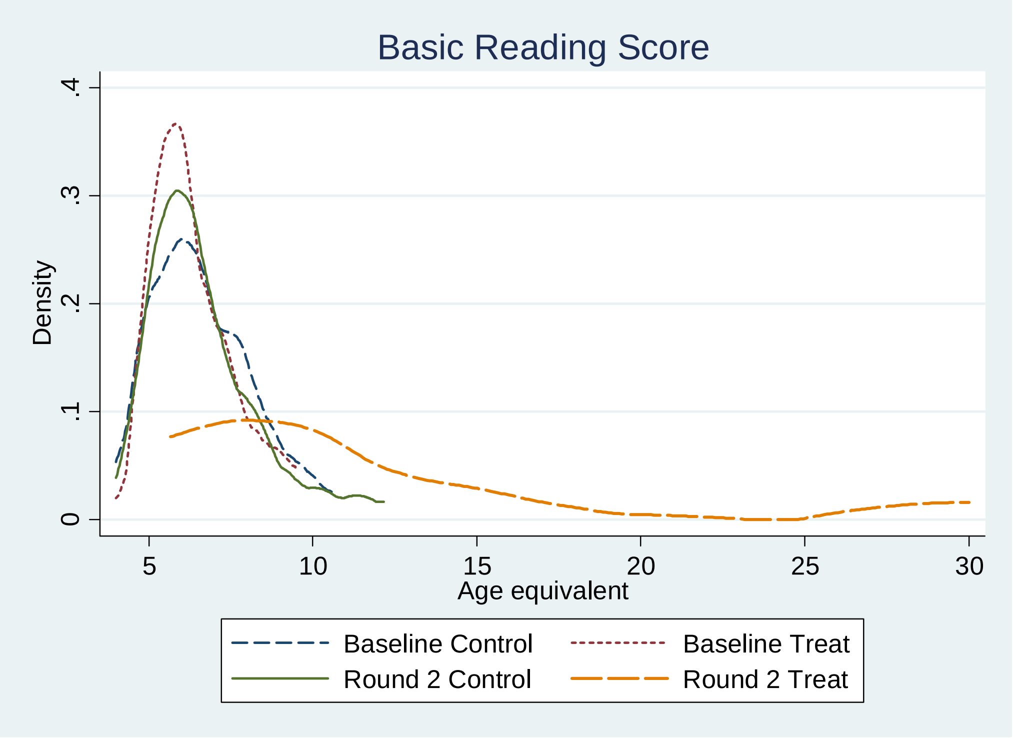


Figure 2C: Distribution of Broad Reading Score by round and treatment status

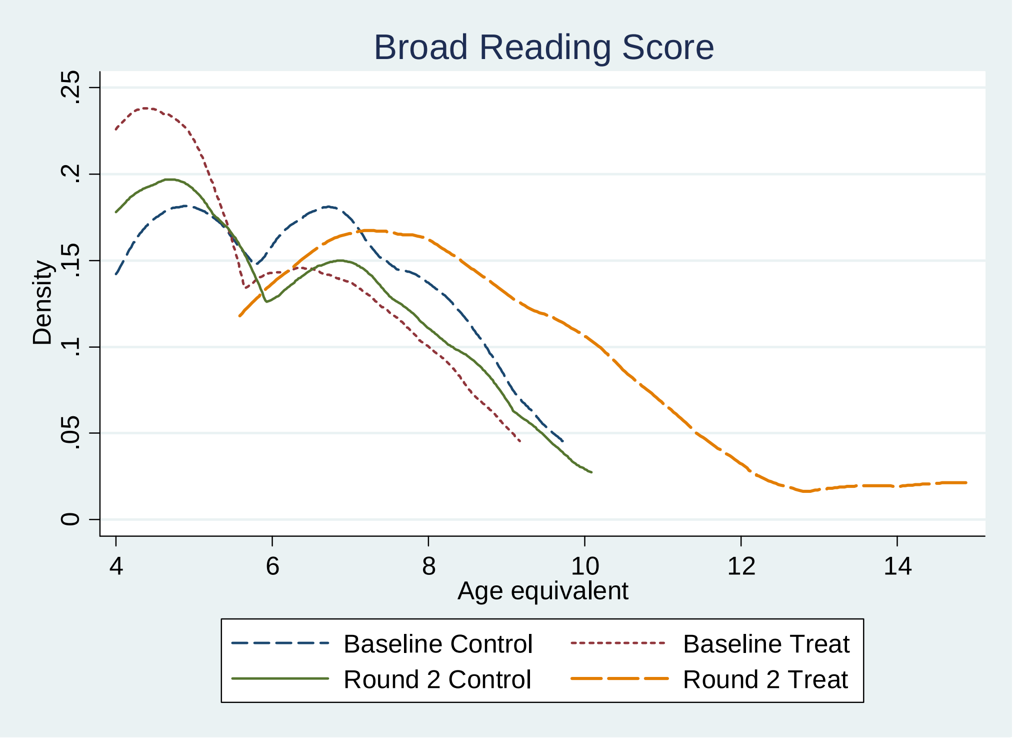


Figure 3A. Phase in Design and Respondents

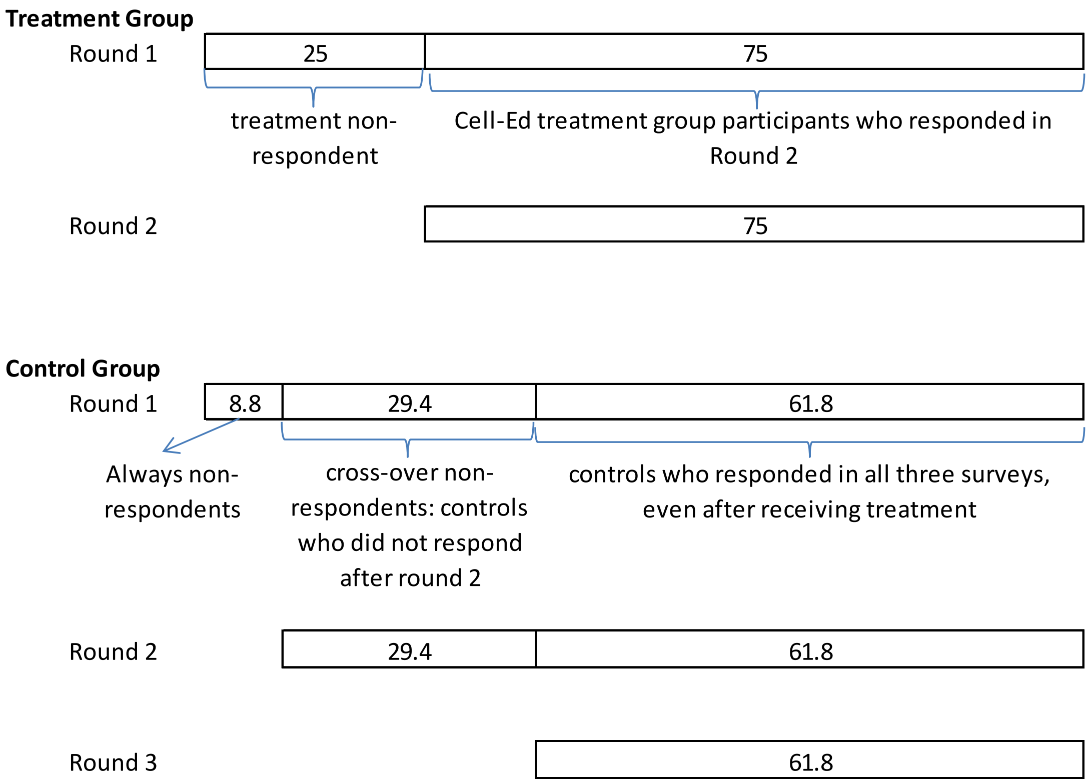

Treatment Group
Round 1
25
75
treatment non-respondent
Cell-Ed treatment group participants who responded in Round 2
Round 2
75
Control Group
Round 1
8.8
29.4
61.8
Always non-respondents
cross-over non-respondents: controls who did not respond after round 2
controls who responded in all three surveys, even after receiving treatment
Round 2
29.4
61.8
Round 3
61.8

Figure 3B. Traditional Lee Bounds and attrition

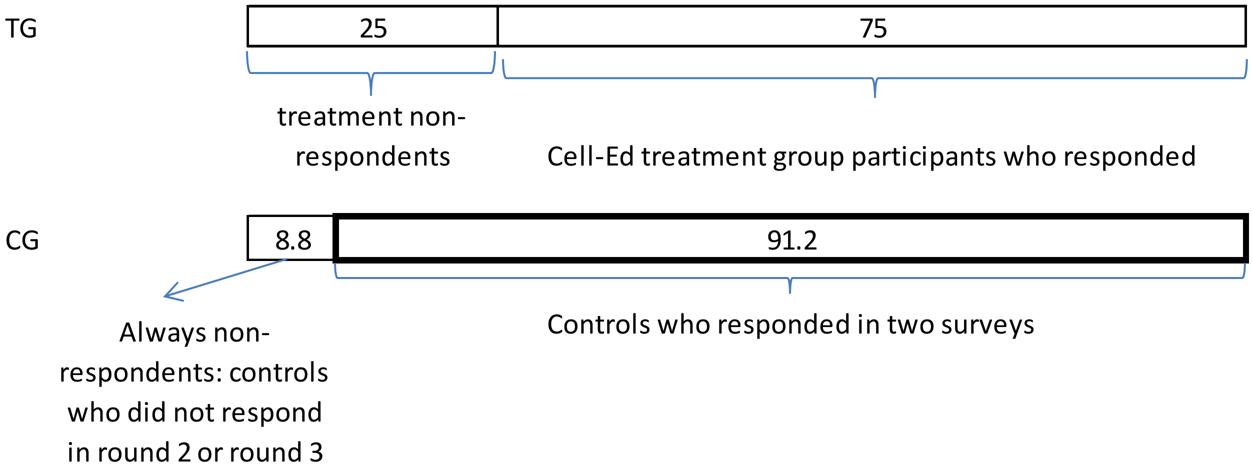
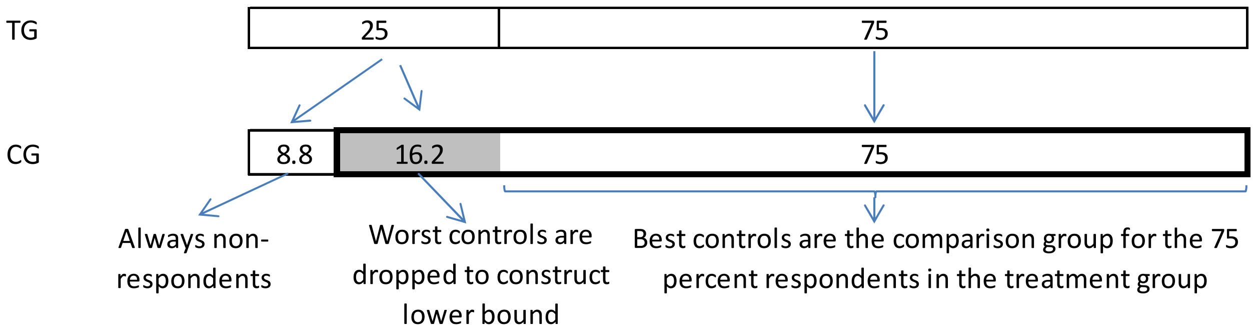
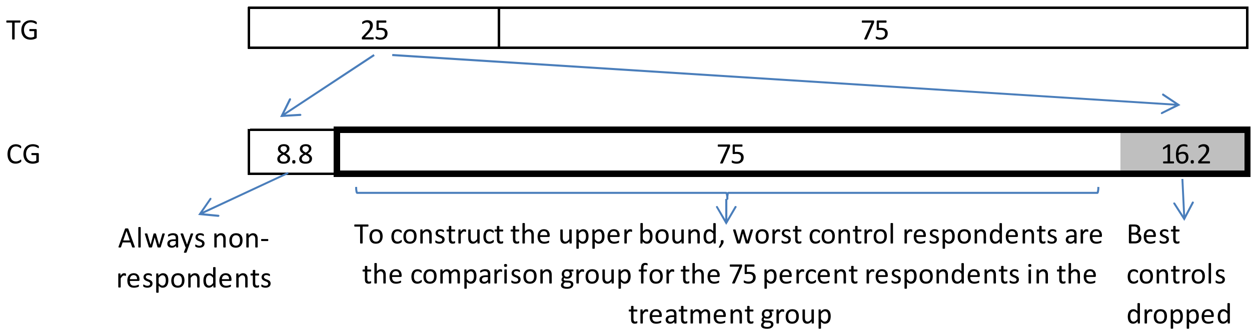

TG
25
75
treatment non-respondents
Cell-Ed treatment group participants who responded
CG
8.8
91.2
Always non-respondents: controls who did not respond in round 2 or round 3
Controls who responded in two surveys

Figure 3C. Traditional Lee Lower Bounds Computation


TG
25
75
CG
8.8
16.2
75
Always non-respondents
Worst controls are dropped to construct lower bound
Best controls are the comparison group for the 75 percent respondents in the treatment group

Notes: Thick black outline indicates control respondents (some of whom are trimmed). The shaded area indicates observations that are trimmed.

Figure 3D. Traditional Lee Upper Bounds Computation


TG
25
75
CG
8.8
75
16.2
Always non-respondents
To construct the upper bound, worst control respondents are the comparison group for the 75 percent respondents in the treatment group
Best controls dropped

Notes: Thick black outline indicates control respondents, some of whom are trimmed. The shaded area indicates observations that are trimmed.

Figure 3E. Adjusted Lee Bounds: respondents and non-respondents

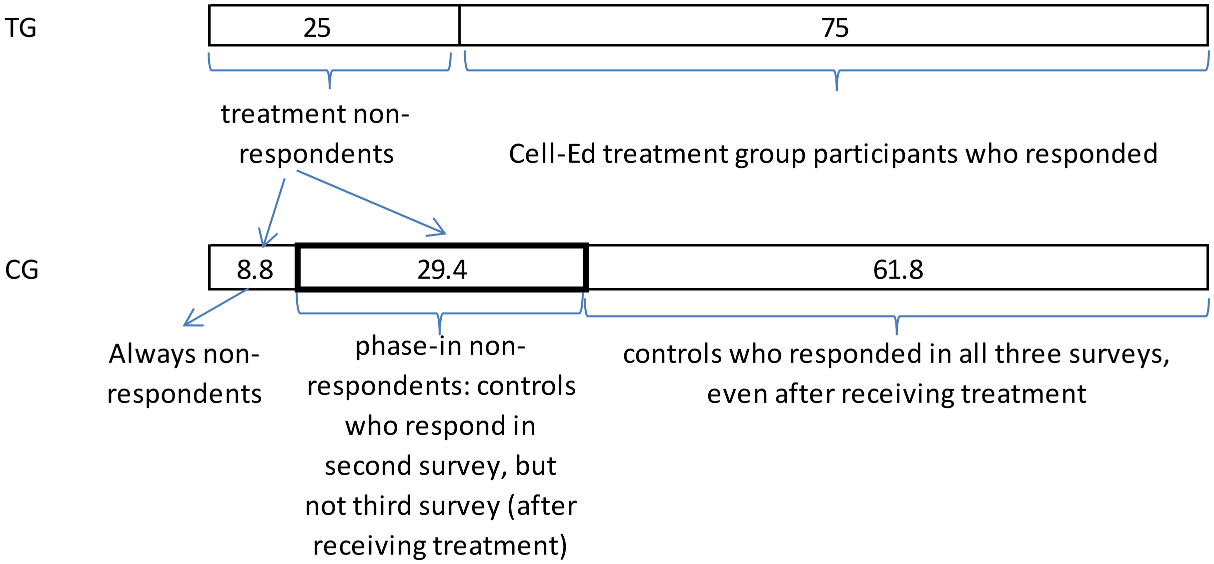
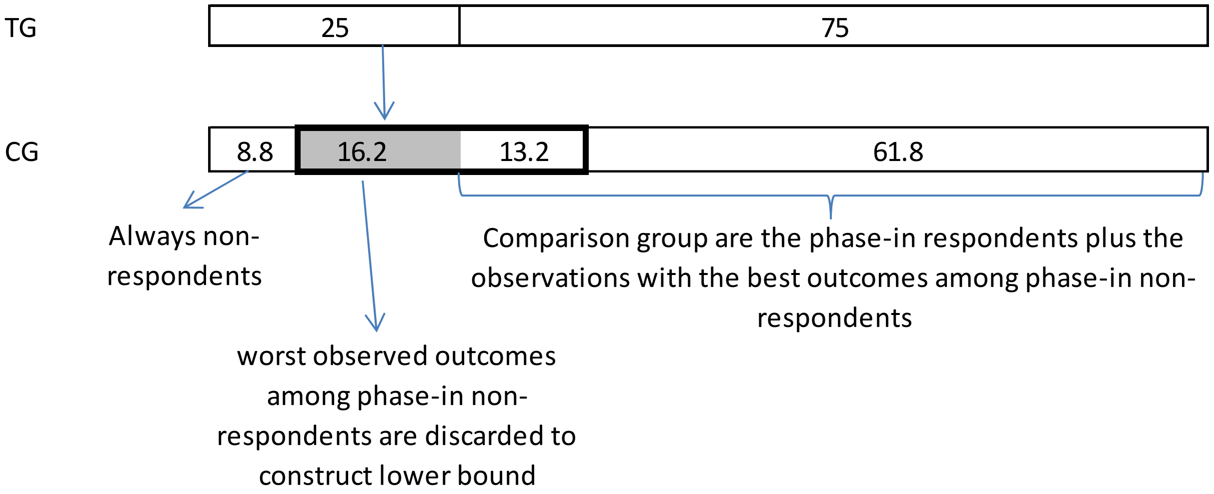
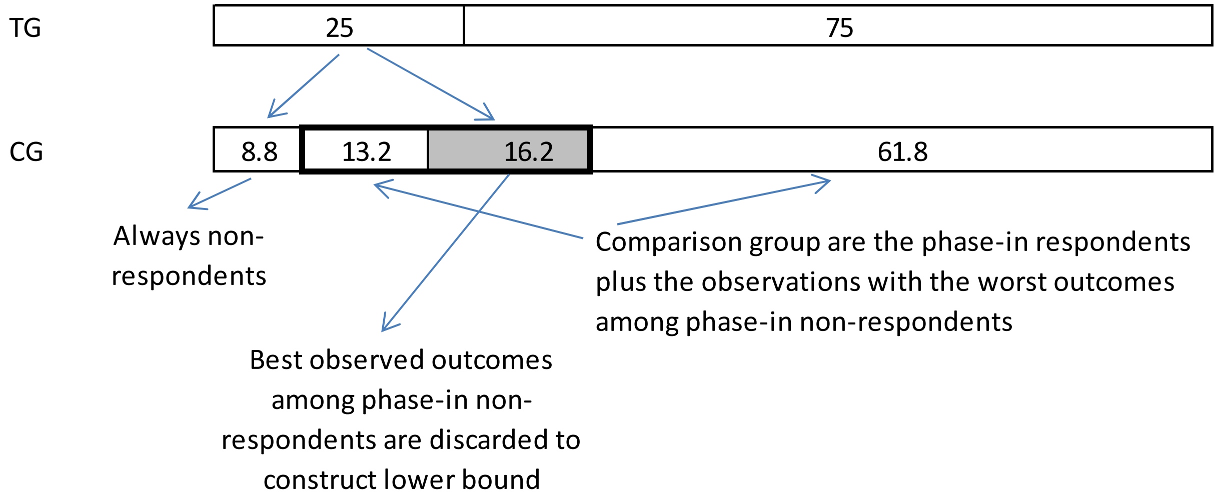

TG
25
75
treatment non-respondents
Cell-Ed treatment group participants who responded
CG
8.8
29.4
61.8
Always non-respondents
phase-in non-respondents: controls who respond in second survey, but not third survey (after receiving treatment)
controls who responded in all three surveys, even after receiving treatment

Figure 3F. Lee Lower Bound using the Phase-in Non-Respondents


TG
25
75
CG
8.8
16.2
13.2
61.8
Always non-respondents
Comparison group are the phase-in respondents plus the observations with the best outcomes among phase-in non-respondents
worst observed outcomes among phase-in non-respondents are discarded to construct lower bound

Notes: Thick black outline indicates control observations who did not respond after receiving treatment (phase-in non-respondents). This is the group of observations from which some trimmed. The shaded area indicates those observations which are trimmed.

Figure 3G. Lee Upper Bound using the Phase-in Non-Respondents


TG
25
75
CG
8.8
13.2
16.2
61.8
Always non-respondents
Comparison group are the phase-in respondents plus the observations with the worst outcomes among phase-in non-respondents
Best observed outcomes among phase-in non-respondents are discarded to construct lower bound

Notes: Thick black outline indicates control observations who did not respond after receiving treatment (phase-in non-respondents). This is the group of observations from which some trimmed. The shaded area indicates those observations which are trimmed.

Figure **4A: When Learners Learn: Call Start and End Times**

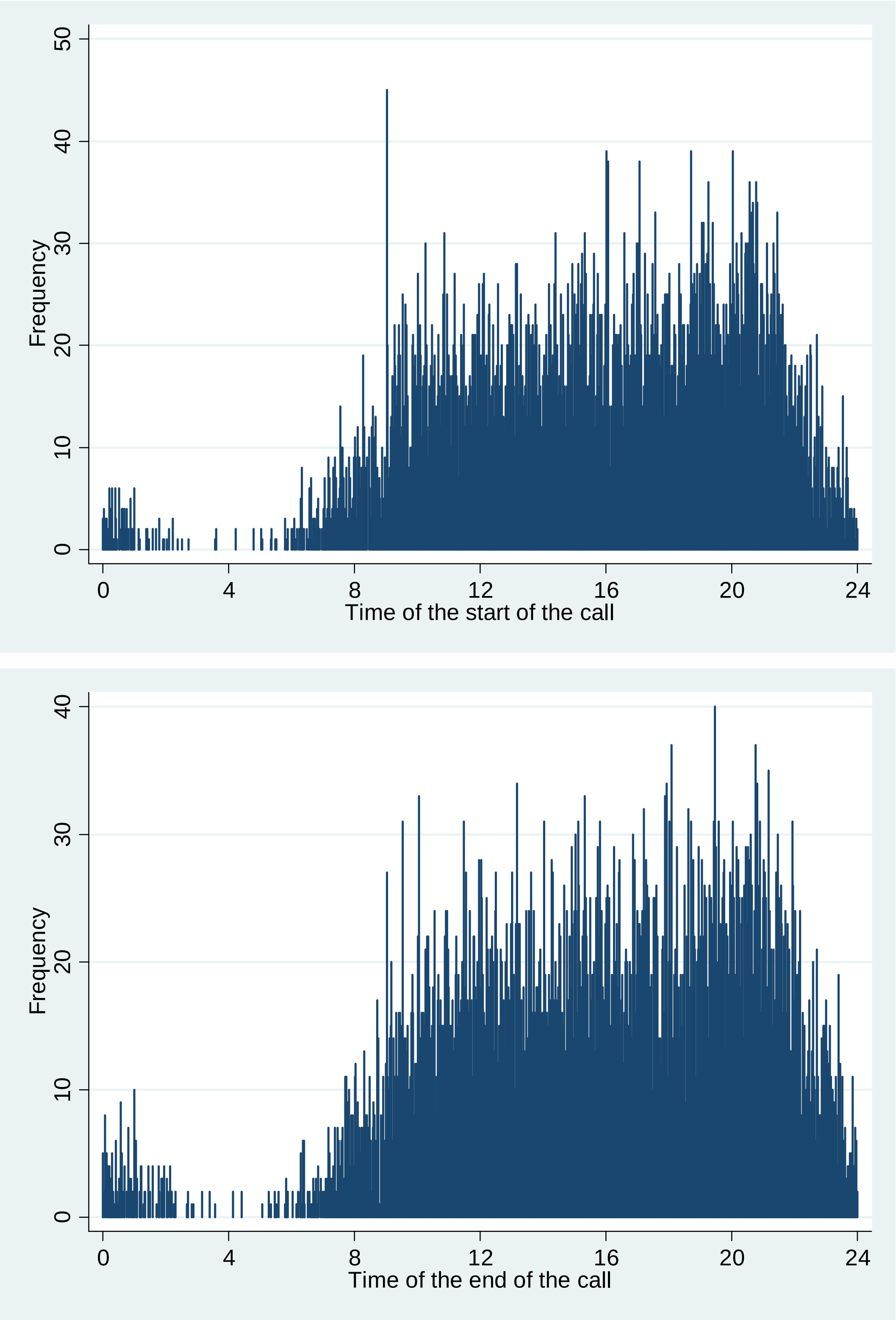


Figure 4B: When Learners Learn: Day of the Week

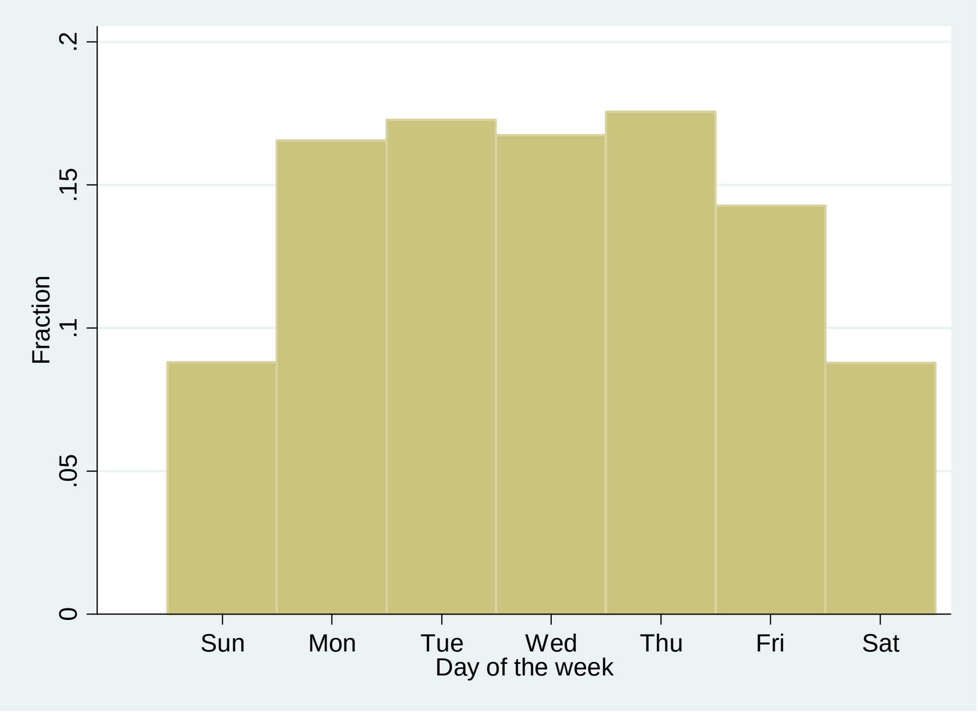


Figure 4C: Duration of Calls in Hours

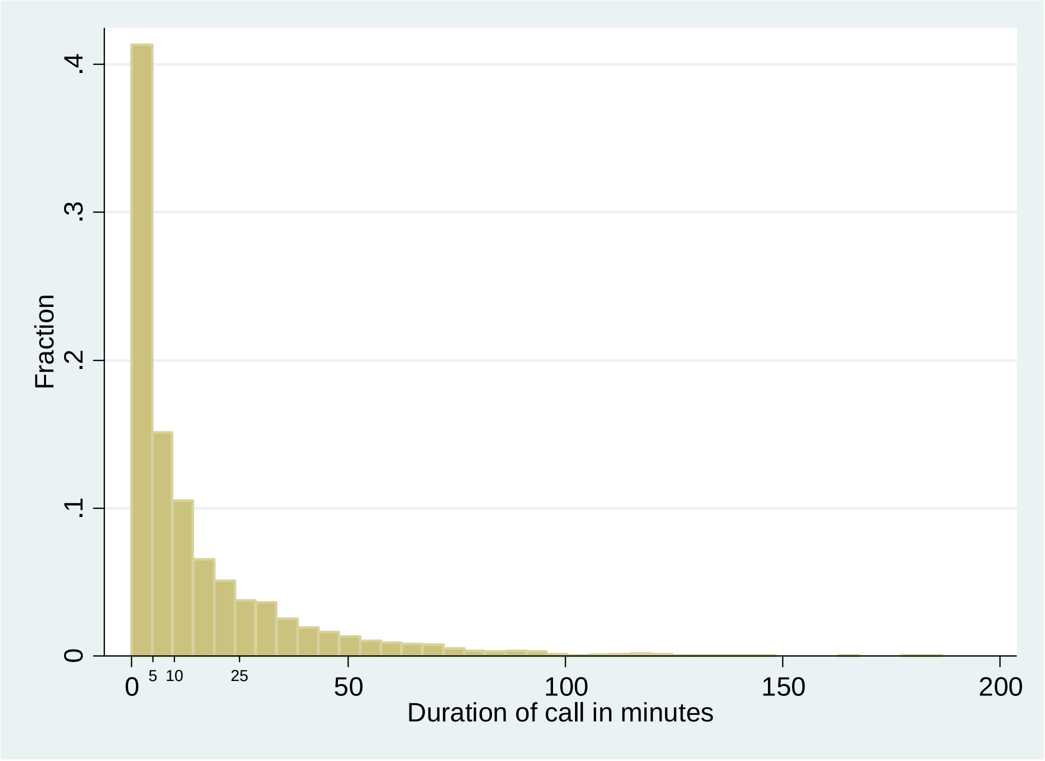


Figure 5A: Time of calls, by employment status

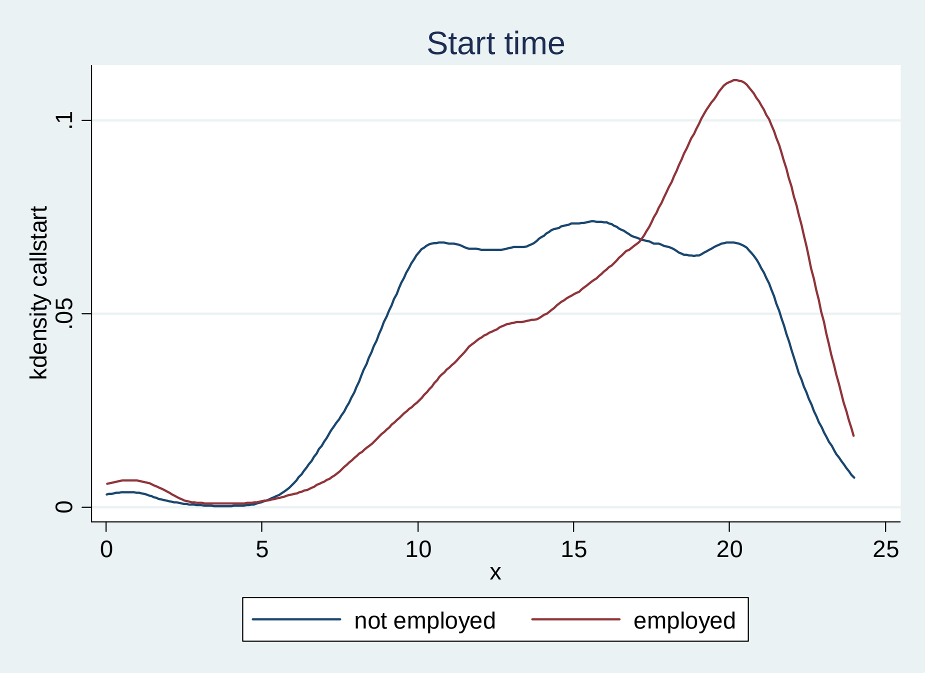


Figure 5B: Day of week for calls, by employment status

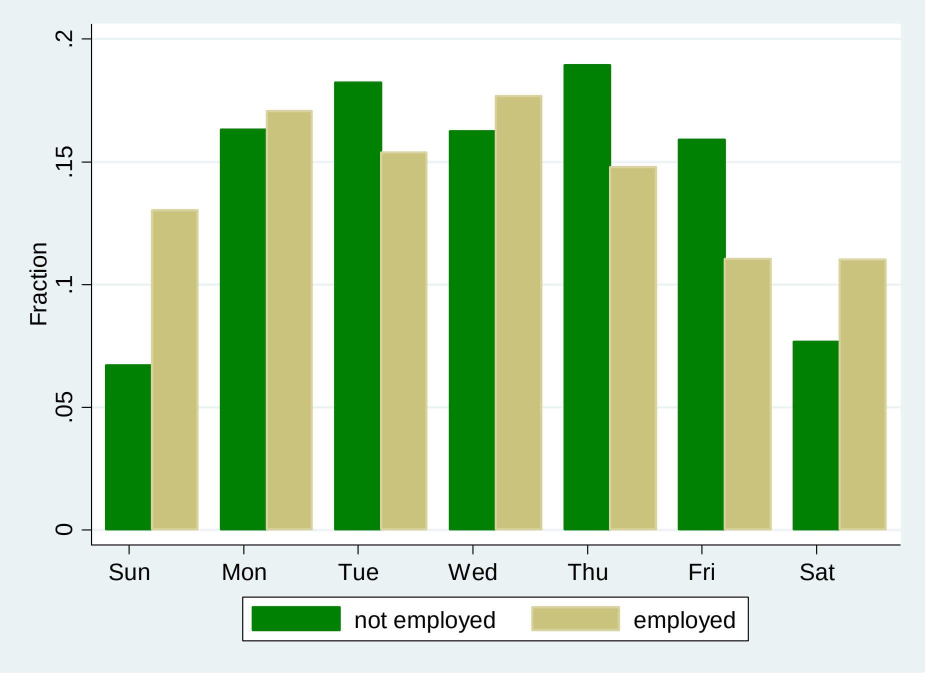


Figure 5C: Call duration in minutes, by employment status

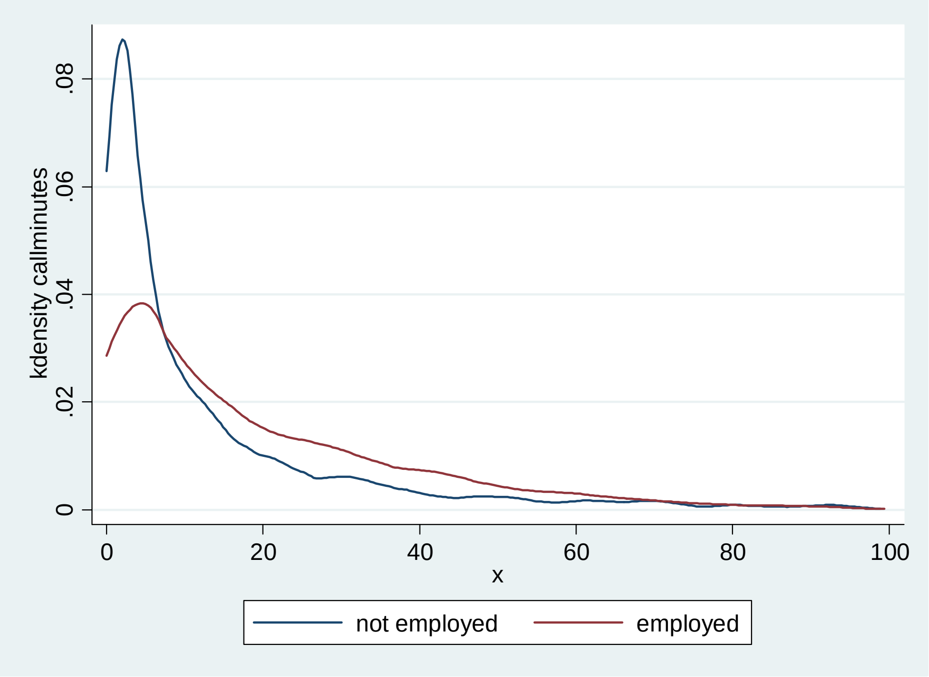


Table 1: Baseline Balance

	Cell-Ed	Initial Control	
	Mean	**Mean**	**Difference**
	(s.d.)	**(s.d.)**	**Coeff (s.e.)**
	(1)	(2)	(3)
Panel A: Covariates			
Age in years	48.50	46.70	1.81
	(12.50)	(13.13)	(3.06)
Male	0.28	0.21	0.07
	(0.45)	(0.41)	(-0.10)
Verbal IQ test	7.74	8.02	-0.28
	(1.85)	(2.31)	(0.50)
Have you ever attended formal school?	0.47	0.56	-0.09
	(0.51)	(0.50)	(0.12)
For how many years did you attend school?	0.80	1.66	-0.85
	(1.10)	(3.18)	(0.58)
Are you currently employed?	0.39	0.35	0.04
	(0.49)	(0.49)	(0.12)
Do you currently own a cell phone?	0.61	0.41	0.20*
	(0.49)	(0.50)	(0.12)
In a normal day, do you make cell phone calls?	0.64	0.67	-0.03
	(0.49)	(0.48)	(0.12)
Panel B: Baseline Outcomes			
Basic reading test, age equivalent	6.36	6.60	-0.24
	(1.27)	(1.55)	(0.34)
Broad reading test, age equivalent	5.56	6.16	-0.60
	(1.66)	(1.79)	(0.42)
Rosenberg Self-Esteem Scale	19.52	18.03	1.49*
	(3.61)	(2.52)	(0.76)
General Self-Efficacy Score	32.88	35.03	-2.15
	(5.97)	(6.05)	(1.47)
Number of observations	36	34	70

Notes: Column 1 presents the mean (and standard deviation) of baseline characteristics for the Cell-Ed treatment group, Column 2 presents the mean (and standard deviation) of baseline characteristics for the control group. Column 3 reports the difference between the two. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table 2A: Survey Non-response

	Cell-Ed		Control		
	(1)	(2)	(3)	(4)	(5)
Round 1: Baseline	36	100%	34	100%	
Round 2: Endline for treatment group	27	75%	31	91%	100%
Round 2: Endline for control/phase-in treated	NA		21	62%	68%

Notes: The rows represent the three survey rounds. Column 1 presents the number of observations for the Cell-Ed treatment group, Column 2 the percentage of treatment observations in that survey round relative to the baseline number (36). Column 3 presents the number of observations for the control group (and later phase-in group). Columns 4 contains the percent control group members observed relative to baseline (34) Column 5 contains the percent observations surveyed relative to the Round 2 (which can be thought of as a second baseline for the phase-in control group), namely 31.

Table 2B: Overview of Interaction with Platform

	Respondents				Non-Respondents			
	mean	sd	min	max	mean	sd	min	max
Panel A: Treatment Group								
Days between first contact and last contact with platform	123.78	29.39	48	150	92.78	45.65	41	150
(minutes)	3144.44	2194.56	469	8886	466.56	515.92	100	1695
Modules completed	401.56	74.48	184	436	31.11	29.16	4	104
Number of observations	27				9			
Panel B: Control/phase-in treatment								
Days between first contact and last contact with platform	133.71	25.47	54	150	103.90	52.71	20	150
(minutes)	2041.14	1183.73	432	3780	707.30	851.19	62	2594
Modules completed	328.33	123.75	104	436	81.40	93.29	2	265
Number of observations	21				10			

Notes: Table presents statistics related to interactions of learners with the Cell-Ed platform. Panel A has outcomes for phase 1 when the treatment group was treated. Panel B has outcomes for phase 2 when the control group was treated (this excludes the 3 observations who were never put on the platform). The four left columns provide information (mean, standard deviation, minimum, maximum) for respondents and the four right columns for non-respondents. Days between first and last contact with the platform is the number of days between the first and the last interaction (+1), which is censured for research purposes above at 150 (in practice, participants could continue on the platform after the respective endline survey). Time spent on the platform are the toal hours spent on the Cell-Ed platform. Modules completed are the total number of modules completed out of 436.

Table 3: OLS Regression Results (Assuming Observations Missing at Random)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Basic Reading					Broad Reading				
	simple	incl	value-			simple	incl	value-		
	means	covariates	added	DD	FE	means	covariates	added	DD	FE
Treatment * Post				4.960***	4.763***				3.003***	2.767***
				(1.434)	(1.383)				(0.681)	(0.452)
Treatment group	5.653***	5.539***	5.840***	-0.024		2.582***	2.533***	3.096***	-0.757**	
	(1.854)	(1.789)	(1.750)	(0.477)		(0.679)	(0.641)	(0.556)	(0.373)	
Post				0.150	0.126				-0.212	-0.177
				(0.465)	(0.145)				(0.442)	(0.132)
Y (t-1)			1.146***					0.909***		
			(0.303)					(0.122)		
Age in years		0.003	-0.019	0.006			0.009	-0.016	0.010	
		(0.047)	(0.042)	(0.024)			(0.024)	(0.013)	(0.014)	
Male		0.524	0.112	0.465			0.243	-0.324	0.367	
		(1.862)	(1.744)	(0.932)			(0.847)	(0.558)	(0.489)	
Verbal IQ test		0.737*	0.562	0.428**			0.326*	0.112	0.283***	
		(0.387)	(0.348)	(0.186)			(0.174)	(0.109)	(0.099)	
Constant	5.214***	-1.388	-7.499	2.548		6.306***	2.974	-0.500	4.165***	
	(1.876)	(5.045)	(4.878)	(2.300)		(0.842)	(2.345)	(1.705)	(1.340)	
Observations	58	58	58	128	128	58	58	57	127	127
R-squared	0.275	0.345	0.415	0.321	0.322	0.315	0.384	0.706	0.363	0.558

Notes: Results from a regression of observed reading test outcomes on different sets of covariates for basic reading scores (Columns 1-5) and broad reading scores (Columns 6-10). Columns 1 and 6 provide estimates from a regression just on the treatment dummy and strata fixed effects. Columns 2 and 7 include covariates (age, gender and IQ). Columns 3 and 8 include the baseline outcome variable denoted by Y(t-1), which is baseline basic reading score for Column 3 and the baseline broad reading score for Column 8. Columns 4 and 9 provide difference-in-difference specifications. Note that the number of observations for the DD specifications (Columns 4 and 9) are higher, as the baseline scores of later control and treatment non-respondents are included. In the value-added and DD specifications for the broad reading score, one observation is missing as the interviewer erroneously stopped the tests after the components of the basic reading score had been completed. Columns 5 and 10 present results from a fixed effects specification. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table 4: Lee Bounds and Adjusted Lee Bounds for the Treatment Effects

	(1)	(2)	(3)	(4)	(5)	(6)
	Basic Reading			Broad Reading		
	Point estimate	Confidence interval lower bound	Confidence interval upper bound	Point estimate	Confidence interval lower bound	Confidence interval upper bound
Panel A: Lee Bounds						
Lower bound	4.433***	**1.583**	7.280	2.079***	**0.723**	3.143
Upper bound	5.437***	2.604	**8.250**	3.116***	1.824	**4.408**
Panel B: Lee Bounds adjusted for phase-in						
Lower bound	4.540***	**1.732**	7.347	2.039***	**0.835**	3.244
Upper bound	4.916***	2.098	**7.734**	2.586***	1.366	**3.805**
Observations	70			70		

Notes: Table presents non-parametric bounds, based on Lee (2009) for basic reading and broad reading. Lee bounds adjusted for phase-in are the bounds that condition on whether the observation was observed in round 2 (for treatment) and round 3 for the control group. We present estimates for the lower and the upper bound. Columns 1 and 4 contain the point estimate for basic and broad reading scores, respectively. Columns 2 and 5 contain the lower bound of the 95% confidence interval and columns 3 and 6 the upper bound of the 95% confidence interval for basic and broad reading scores, respectively. All confidence intervals are bootstrapped. The lower 95% CI bound and the upper 95% CI bound are overly conservative bounds on the treatment effect (as Imbens and Manski (2004) show), so are wider than the 95% confidence interval for the treatment effect. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table 5: Non-parametric Lower Bounds on the Treatment Effect Implementing Monotone Treatment Response and Monotone Treatment Selection Assumptions

	(1)	(2)	(3)	(4)	(5)	(6)
	Basic Reading			Broad Reading		
	Point estimate	Confidence interval lower bound	Confidence interval upper bound	Point estimate	Confidence interval lower bound	Confidence interval upper bound
Lower bound	3.310***	**1.068**	5.551	1.540***	**0.477**	2.602
Observations	70			70		

Notes: Table presents non-parametric lower bounds the average treatment effect, based on implementing monotone treatment response and monotone treatment selection assumptions for basic reading and broad reading. Columns 1 and 4 contain the point estimate for basic and broad reading scores, respectively. Columns 2 and 5 contain the lower bound of the 95% confidence interval and columns 3 and 6 the upper bound of the 95% confidence interval for basic and broad reading scores, respectively. All confidence intervals are bootstrapped. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table 6A: OLS Results for Empowerment (Assuming Observations Missing at Random)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Rosenberg Self-Esteem Score					General Self-Efficacy Scale				
	simple	incl	value-			simple	incl	value-		
	means	covariates	added	DD	FE	means	covariates	added	DD	FE
Treatment * Post				1.115	1.379				3.747*	4.757***
				(1.270)	(1.395)				(2.203)	(1.693)
Treatment group	2.490**	2.572**	2.601**	1.115		1.422	1.325	3.132	-2.264	
	(1.099)	(1.053)	(1.235)	(0.766)		(2.103)	(1.912)	(2.263)	(1.474)	
Post				1.748**	1.621**				-0.656	-0.966
				(0.792)	(0.757)				(1.664)	(1.357)
Y (t-1)			-0.060					0.463**		
			(0.268)					(0.194)		
Age in years		0.040	0.038	0.002			0.092	0.051	0.090*	
		(0.045)	(0.049)	(0.025)			(0.065)	(0.071)	(0.049)	
Male		0.810	0.893	1.333*			0.442	0.173	0.001	
		(1.246)	(1.509)	(0.735)			(2.061)	(2.043)	(1.318)	
Verbal IQ test		0.494	0.512	0.335*			-0.053	-0.213	0.055	
		(0.351)	(0.384)	(0.186)			(0.616)	(0.564)	(0.343)	
Constant	22.071***	14.832***	15.993**	16.628***		35.756***	31.043***	17.943**	30.210***	
	(1.746)	(4.693)	(6.826)	(2.422)		(2.377)	(7.052)	(7.015)	(4.139)	
Observations	55	55	53	122	122	55	55	53	122	122
R-squared	0.224	0.285	0.284	0.294	0.192	0.068	0.103	0.296	0.099	0.152

Notes: Results from a regression of observed reading test outcomes on different sets of covariates for the Rosenberg Self-esteem score (Columns 1-5) and the General self-efficacy Score (Columns 6-10). Columns 1 and 6 provide estimates from a regression just on the treatment dummy and and strata fixed effects. Columns 2 and 7 include covariates (age, gender and IQ). Columns 3 and 8 include the baseline outcome variable denoted by Y(t-1), which is baseline self-esteem score for Column 3 and the self-efficacy score for Column 8. Columns 4 and 9 provide difference-in-difference specifications. Note that the number of observations for the DD specifications (Columns 4 and 9) are higher, as the baseline scores of later control and treatment non-respondents are included. Columns 5 and 10 present results from a fixed effects specification. Relative to Table 3, 6 observations are missing for the DD specifications, as some respondents could not understand and could not complete the measures for self-esteem and self-efficacy. This also reduces the mean difference specifications and the value-added specifications. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table 6B: Lee Bounds and Adjusted Lee Bounds for the Treatment Effects on Measures of Empowerment

	(1)	(2)	(3)	(4)	(5)	(6)
	Rosenberg Self-Esteem Score			General Self-Efficacy Scale		
	Point estimate	Confidence interval lower bound	Confidence interval upper bound	Point estimate	Confidence interval lower bound	Confidence interval upper bound
Panel A: Lee Bounds						
Lower bound	1.80	**-0.644**	4.236	-0.90	**-3.996**	2.191
Upper bound	3.71***	1.190	**6.230**	2.56	-1.271	**6.389**
Panel B: Lee Bounds adjusted for phase-in						
Lower bound	2.27*	**-0.209**	4.740	1.23	**-2.722**	5.190
Upper bound	3.13**	0.689	**5.570**	2.10	-1.775	**5.971**
Observations	70			70		

Notes: Table presents non-parametric bounds, based on Lee (2009) for self-esteem and self-efficacy. Lee bounds adjusted for phase-in are the bounds that condition on whether the observation was observed in round 2 (for treatment) and round 3 for the control group. We present estimates for the lower and the upper bound. Columns 1 and 4 contain the point estimate forself-esteem and self-efficacy, respectively. Columns 2 and 5 contain the lower bound of the 95% confidence interval and columns 3 and 6 the upper bound of the 95% confidence interval for self-esteem and self-efficacy, respectively. All confidence intervals are bootstrapped. The lower 95% CI bound and the upper 95% CI bound are overly conservative bounds on the treatment effect (as Imbens and Manski (2004) show), so are wider than the 95% confidence interval for the treatment effect. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively. The lower bound in the phase-in design is bounds the treatment away from zero at the 10 percent level of significance, but not (as column (2) shows) at the 5 percent level.

Table 6C: Non-parametric Lower Bounds on the Treatment Effect Implementing Monotone Treatment Response and Monotone Treatment Selection Assumptions - Measures of Empowerment

	(1)	(2)	(3)	(4)	(5)	(6)
	Rosenberg Self-Esteem Score			General Self-Efficacy Scale		
	Point estimate	Confidence interval lower bound	Confidence interval upper bound	Point estimate	Confidence interval lower bound	Confidence interval upper bound
Lower bound	2.303**	**0.397**	4.209	1.366	**-1.455**	4.186
Observations	69			69		

Notes: Table presents non-parametric lower bounds the average treatment effect, based on implementing monotone treatment response and monotone treatment selection assumptions for self-esteem and self-efficacy. Columns 1 and 4 contain the point estimate for self-esteem and self-efficacy, respectively. Columns 2 and 5 contain the lower bound of the 95% confidence interval and columns 3 and 6 the upper bound of the 95% confidence interval for self-esteem and self-efficacy, respectively. All confidence intervals are bootstrapped. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively. One observation in the treatment group who otherwise completed the tests did not understand the questions neither in baseline nor endline and was dropped.

Table 7: Determinants of empowerment

	(1)	(2)	(3)	(4)	(5)
	Rosenberg Self-Esteem 4-item sub-scale				
	proportion incorrect	only females	only males	gender interaction	fixed effects
Proportion incorrect	-0.963*	-0.0708	-1.657**	-0.209	0.279
	(0.519)	(0.486)	(0.586)	(0.535)	(0.306)
Male * Proportion incorrect				-1.799**	-0.895*
				(0.870)	(0.518)
Age in years	0.0103	-0.00539	0.00716	0.0112	
	(0.0133)	(0.0196)	(0.0221)	(0.0129)	
Male	0.101			0.110	
	(0.456)			(0.435)	
Verbal IQ test	0.238*	0.224	0.0484	0.234*	
	(0.455)			(0.436)	
Constant	6.574***	7.077***	8.610***	6.329***	8.848***
	(1.180)	(1.461)	(1.743)	(1.137)	(0.109)
Observations	638	411	227	638	638
R-squared	0.096	0.077	0.318	0.111	0.594

Notes: Results from a regression of 4-item subset of Rosenberg Self-esteem scale on different sets of covariates, using the weekly surveys with respondents. Column 1 contains the basic specification that includes baseline covariates and the proportion of errors made in the previous week on the platform. Columns 2 and 3 present results for the subsample of females and males, respectively. Column 4 includes the proportion of errors made in interactions with the platform, as well as an interaction with gender. Column 5 provides a fixed effects specification. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A1: Baseline Comparisons

	Whole sample			Round 1 Respondents		
	Respondent	**Non-respondent**		**Cell-Ed**	**Initial Control**	
	Mean	**Mean**	**Difference**	**Mean**	**Mean**	**Difference**
	(s.d.)	**(s.d.)**	**Coeff (s.e.)**	**(s.d.)**	**(s.d.)**	**Coeff (s.e.)**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Covariates						
Age in years	46.74	49.57	-2.84	46.43	47.49	-1.06
	(12.95)	(12.37)	(3.29)	(13.69)	(12.99)	(3.52)
Male	0.29	0.14	0.16	.19	.37	-.18
	(0.46)	(0.35)	(0.11)	(.4)	(.49)	(.12)
Verbal IQ test	7.90	7.81	0.09	7.89	7.77	.12
	(2.05)	(2.18)	(0.54)	(2.34)	(1.98)	(.57)
Have you ever attended formal school?	0.44	0.68	-0.24*	.55	.44	.1
	(0.5)	(0.48)	(0.13)	(.51)	(.51)	(.13)
For how many years did you attend school?	1.11	1.46	-0.35	1.59	.85	.74
	2.39)	(2.36)	(0.64)	(3.27)	(1.2)	(.68)
Are you currently employed?	0.33	0.45	-0.12	.35	.41	-.05
	(0.48)	(0.51)	(0.13)	(.49)	(.5)	(.13)
Do you currently own a cell phone?	0.54	0.45	0.09	.45	.63	-.18
	(0.50)	(0.51)	(0.13)	(.51)	(.49)	(.13)
In a normal day, do you make cell phone calls	0.69	0.57	0.12	.68	.7	-.03
	(0.47)	(0.51)	(0.13)	(.48)	(.47)	(.12)
Panel B: Baseline Outcomes						
Basic reading test, age equivalent	6.61	6.20	0.41	6.54	6.57	-.03
	(1.51)	(1.12)	(0.36)	(1.61)	(1.34)	(.39)
Broad reading test, age equivalent	6.01	5.52	0.49	6.08	5.75	.33
	(1.80)	(1.58)	(0.45)	(1.85)	(1.76)	(.48)
Rosenberg Self-Esteem Scale	18.78	18.73	0.05	18.13	19.58	-1.45*
	(3.04)	(3.48)	(0.83)	(2.6)	(3.24)	(.79)
General Self-Efficacy Score	33.09	35.77	-2.68*	35.42	32.29	3.13*
	(6.45)	(4.83)	(1.55)	(6.13)	(6.1)	(1.66)
Number of observations	48	22	70	27	31	58

Notes: Column 1 presents the mean (and standard error) of baseline characteristics for the observations that responded in round 2 for the treatment group and in round 3 for the treatment and control group, Column 2 presents the mean (and standard error) for participants that did not respond. Column 3 reports the Difference between the two. Clumns 4-5 present differences between the treatment respondents and the control respondents in Round 1. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A2: Baseline Determinants of Non-Response (Logistic Regression)

Dep Var. Non-Response	(1)	(2)	(3)	(4)	(5)
Age in years	0.005	0.009*	0.008	0.010*	0.010*
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Male	-0.209*	-0.229**	-0.229**	-0.204*	-0.259**
	(0.113)	(0.109)	(0.106)	(0.123)	(0.110)
Verbal IQ test	0.008	0.006	0.005	0.015	-0.005
	(0.028)	(0.030)	(0.029)	(0.031)	(0.034)
Have you ever attended formal school?		0.278**	0.335***	0.251**	0.359***
		(0.119)	(0.119)	(0.121)	(0.129)
For how many years did you attend school?		-0.007	-0.020		-0.031
		(0.029)	(0.029)		(0.030)
Are you currently employed?		0.175	0.244*		0.263
		(0.134)	(0.142)		(0.163)
Do you currently own a cell phone?			-0.146		-0.346*
			(0.143)		(0.192)
In a normal day, do you make cell phone calls?			-0.124		-0.031
			(0.155)		(0.182)
Basic reading test, age equivalent				-0.065	-0.092
				(0.114)	(0.134)
Broad reading test, age equivalent				-0.003	0.037
				(0.088)	(0.106)
Rosenberg Self-Esteem Scale				0.009	0.032
				(0.021)	(0.024)
General Self-Efficacy Score				0.013	0.024*
				(0.011)	(0.014)
Observations	70	66	65	67	62
Pseudo R-squared	0.0398	0.135	0.204	0.140	0.316

Notes: Table presents results from a logistic regression of non-response on baseline covariates. Non-response=1 is defined as not responding in round 2 for the treatment group and not responding in round 2 or 3 for the control group. Marginal effects are reported. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A3 : Outcomes Pre and Post

	Cell-Ed	**Initial Control**		
	Mean	**Mean**	**Difference**	**Difference**
	(s.d.)	**(s.d.)**	**Coeff (s.e.)**	**Coeff (s.e.)**
	(1)	(2)	(3)	(4)
Panel A: Basic Reading				
Baseline	6.36	6.60	-0.24	
	(1.27)	(1.55)	(0.34)	
Round 2 (Endline for the treatment group)	11.46	6.67	4.79 ***	
	(7.37)	(1.72)	(1.36)	
Round 3 (Endline for the phase-in group)		11.42		0.04
		(7.86)		(2.21)
Panel B: Broad Reading				
Baseline	5.56	6.16	-0.60	
	(1.66)	(1.79)	(0.42)	
Round 2 (Endline for the treatment group)	8.39	5.90	2.49 ***	
	(2.49)	(1.90)	(0.58)	
Round 3 (Endline for the phase-in group)		8.19		0.20
		(3.41)		(0.85)
Panel C: Rosenberg Self-Esteem Scale				
Baseline	19.52	18.03	1.49	
	(3.61)	(2.52)	(0.76) *	
Round 2 (Endline for the treatment group)	22.54	19.76	2.78 **	
	(4.48)	(3.65)	(1.10)	
Round 3 (Endline for the phase-in group)		22.45		0.09
		(5.20)		(1.43)
Panel D: General Self-Efficacy Scale				
Baseline	32.88	35.03	-2.15	
	(5.97)	(6.05)	(1.47)	
Round 2 (Endline for the treatment group)	35.96	34.41	1.55	
	(5.23)	(6.98)	(1.68)	
Round 3 (Endline for the phase-in group)		35.25		0.71
		(5.82)		(1.63)
Number of Observations				
Baseline	36	34	70	
Round 2 (Endline for the treatment group)	27	31	58	
Round 3 (Endline for the phase-in group)		21		

Notes: The four panels contain information on the four outcomes. The rows within the panels are for the three survey rounds. Column 1 presents the mean level for Cell-Ed treatment participants at baseline, Column 2 presents the mean for Cell-Ed control participants. Column 3 reports the within-round difference. Column 4 reports the difference between Round 3 phase-in control outcomes and Round 2 treatment group outcomes. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A4: Robustness checks for simple regression results (assuming observations missing at random)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Basic Reading					Broad Reading				
	basic specification	self-esteem	baseline cell	Z-scores	Z-scores	basic specification	self-esteem	baseline cell	Z-scores	Z-scores
Treatment	5.539***	5.355***	5.959***	3.939***	3.224***	2.533***	2.524***	2.712***	1.457***	1.336***
	(1.413)	(1.549)	(1.400)	(1.005)	(0.822)	(0.606)	(0.671)	(0.602)	(0.348)	(0.319)
Age in years	0.00317	0.0112	-0.0140	0.00225	0.00184	0.00856	0.00864	0.00145	0.00492	0.00451
	(0.0540)	(0.0582)	(0.0535)	(0.0384)	(0.0314)	(0.0231)	(0.0252)	(0.0230)	(0.0133)	(0.0122)
Male	0.524	-0.685	0.179	0.373	0.305	0.243	-0.0492	0.0719	0.140	0.128
	(1.600)	(1.810)	(1.598)	(1.138)	(0.931)	(0.686)	(0.784)	(0.688)	(0.394)	(0.362)
Verbal IQ test	0.737**	0.653*	0.886**	0.524**	0.429**	0.326**	0.315*	0.386**	0.187**	0.172**
	(0.346)	(0.373)	(0.347)	(0.246)	(0.202)	(0.148)	(0.162)	(0.149)	(0.0854)	(0.0783)
Baseline Rosenberg score		0.351					0.0399			
		(0.311)					(0.135)			
Baseline cell ownership			-2.887*					-1.122*		
			(1.476)					(0.635)		
Constant	-1.388	-7.760	0.769	-5.594	-4.690	2.974	2.385	3.894*	-1.657	-1.542
	(4.905)	(7.786)	(4.929)	(3.488)	(2.855)	(2.102)	(3.371)	(2.120)	(1.209)	(1.109)
Observations	58	55	57	58	58	58	55	57	58	58
R-squared	0.345	0.363	0.398	0.345	0.345	0.384	0.388	0.430	0.384	0.384

Notes: Table presents a results from an OLS regression of basic reading scores (Columns 1 - 3) and broad reading scores (Columns 6-8) on covariates that are unbalanced at baseline. For comparison purposes, Columns 1 and 6 contain the results from Columns 2 and 7 of Table 3, respectively. The additional covariates include baseline Rosenberg self-esteem score (in Columns 2 and 7) and baseline cell phone ownership (Columns 3 and 8). The dependent variables in Columns 4 and 9 are Z-scores of basic and broad reading scores, respectively, which are normalized by the respective *baseline* mean score and standard deviation. Columns 5 and 10 contain Z-scores of basic and broad reading scores, respectively, which are normalized by the respective control group mean score and standard deviation from that survey round. Columns 4 - 5 and 9 - 10 control for the same covariates as Columns 1 and 6. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A5: Robustness checks for simple regression results (assuming observations missing at random) for empowerment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Rosenberg Self-Esteem Score					General Self-Efficacy Scale				
	basic specification	self-esteem	baseline cell	Z-scores	Z-scores	basic specification	self-esteem	baseline cell	Z-scores	Z-scores
Treatment	2.572**	2.601*	2.594**	0.812**	0.704**	1.325	1.172	1.145	0.219	0.190
	(1.173)	(1.295)	(1.202)	(0.370)	(0.321)	(1.914)	(1.960)	(1.925)	(0.316)	(0.274)
Age in years	0.0402	0.0380	0.0404	0.0127	0.0110	0.0916	0.134*	0.107	0.0151	0.0131
	(0.0446)	(0.0490)	(0.0461)	(0.0141)	(0.0122)	(0.0728)	(0.0742)	(0.0739)	(0.0120)	(0.0104)
Male	0.810	0.893	0.693	0.256	0.222	0.442	-2.025	0.573	0.0730	0.0634
	(1.273)	(1.469)	(1.332)	(0.402)	(0.349)	(2.078)	(2.224)	(2.134)	(0.343)	(0.298)
Verbal IQ test	0.494	0.512	0.491	0.156	0.135	-0.0526	-0.328	-0.105	-0.00868	-0.00754
	(0.304)	(0.325)	(0.311)	(0.0960)	(0.0833)	(0.496)	(0.493)	(0.498)	(0.0819)	(0.0711)
Baseline Rosenberg score		-0.0596					0.854**			
		(0.251)					(0.380)			
Baseline cell ownership			0.216					2.763		
			(1.254)					(2.009)		
Constant	14.83***	15.99**	14.98***	-1.240	-1.349	31.04***	14.55	28.55***	-0.483	-0.483
	(4.312)	(6.502)	(4.631)	(1.361)	(1.181)	(7.037)	(9.841)	(7.418)	(1.161)	(1.008)
Observations	55	53	54	55	55	55	53	54	55	55
R-squared	0.285	0.284	0.289	0.285	0.285	0.103	0.207	0.141	0.103	0.103

Notes: Table presents a results from an OLS regression of the Rosenberg Self-Esteem Score (Columns 1 - 3) and the General Self-Efficacy Scale Score (Columns 6-8) on covariates that are unbalanced at baseline. For comparison purposes, Columns 1 and 6 contain the results from Columns 2 and 7 of Table 3, respectively. The additional covariates include baseline Rosenberg self-esteem score (in Columns 2 and 7) and baseline cell phone ownership (Columns 3 and 8). (Note that Column 2 is equivalent to a value added specification). The dependent variables in Columns 4 and 9 are Z-scores of the Rosenberg Self-Esteem Score and General Self-Efficacy Scale Score, respectively, which are normalized by the respective baseline mean score and standard deviation. Columns 5 and 10 contain Z-scores of the Rosenberg Self-Esteem Score and General Self-Efficacy Scale Score, respectively, which are normalized by the respective control group mean score and standard deviation from that survey round. Columns 4 - 5 and 9 - 10 control for the same covariates as Columns 1 and 6. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A6: OLS Results for Index of Empowerment (Assuming Observations Missing at Random)

	(1)	(2)	(3)	(4)	(5)	(6)
	Index of Empowerment					
	simple means	incl covariates	value-added	DD	FE	baseline cell
Treatment * Post				0.970	1.220**	
				(0.631)	(0.575)	
Treatment group	1.021	1.031*	1.207*	-0.021		1.008*
	(0.611)	(0.536)	(0.649)	(0.406)		(0.583)
Post				0.444	0.352	
				(0.451)	(0.371)	
baseline empowerment score			0.427*			
			(0.233)			
Age in years		0.028	0.026	0.016		0.0304
		(0.021)	(0.020)	(0.014)		(0.0224)
Male		0.329	-0.044	0.421		0.313
		(0.637)	(0.692)	(0.379)		(0.646)
Verbal IQ test		0.147	0.088	0.115		0.138
		(0.191)	(0.188)	(0.099)		(0.151)
Baseline cell ownership						0.524
						(0.608)
Constant	1.339*	-1.723	-1.297	-1.294	-0.003	-2.089
	(0.788)	(2.290)	(2.075)	(1.238)	(0.129)	(2.246)
Observations	55	55	53	122	122	54
R-squared	0.152	0.203	0.289	0.182	0.223	0.219

Notes: Results from a regression of an index of empowerment observed reading test outcomes on covariates as in Tables 3 and 6A. Column 1 provides estimates from a regression just on the treatment dummy and strata fixed effects. Columns 2 include covariates (age, gender and IQ). Column 3 includes the baseline empowerment score. Column 4 provides a difference-in-difference specification. Note that the number of observations for the DD specification (Column 4) is higher, as the baseline scores of later control and treatment non-respondents are included. Column 5 presents results from a fixed effects specification. Column 6 contains the robustness check which controls for baseline mobile phone ownership. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A7: Results for empowerment (weekly observations)

	(1)	(2)	(3)	(4)	(5)
	Rosenberg Self-Esteem Score (4-item subset)				
	Basic specification	incl covariates	fixed effects	incl covariates	fixed effects
Treatment * Post	0.497**	0.435*	0.459***	0.411**	0.494***
	(0.234)	(0.222)	(0.0994)	(0.188)	(0.0967)
Treatment group	-0.284	-0.154		-0.0918	
	(0.446)	(0.378)		(0.341)	
Age in years		0.000693		-0.000307	
		(0.0113)		(0.0111)	
Male		0.00601		0.0784	
		(0.416)		(0.400)	
Verbal IQ test		0.172		0.162	
		(0.112)		(0.112)	
Constant	8.250***	6.960***	8.569***	7.017***	8.566***
	(0.497)	(1.113)	(0.0703)	(1.081)	(0.0666)
include face to face survey measurements	No	No	No	Yes	Yes
Observations	1,135	1,135	730	1,272	809
R-squared	0.019	0.052	0.530	0.052	0.519

Notes: Results from a regression of a 4-item subset of the Rosenberg Self-Esteem scale and covariates. Sample includes treatment, control and phase-in observations. Column 1 provides estimates from a regression on treatment dummy and a dummy for whether the observation was in the intial treatment or control group. Columns 2 and 4 include covariates (age, gender and IQ). Columns 3 and 5 provide fixed effects specifications including only the phase-in observations. Columns 4 and 5 include the face to fact measurements in addition to weekly telephone measurements. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A8: Learning determinants of non-response (logistic regression)

	(1)	(2)	(3)	(4)	(5)	(6)
	Non-response					
	whole sample	only females	only females	whole sample	only females	only females
Average proportion incorrect is larger than 0.4	0.271*	0.350**	0.341**			
	(0.152)	(0.171)	(0.173)			
Average proportion incorrect				0.399	1.071**	1.046*
				(0.281)	(0.540)	(0.569)
Average proportion incorrect squared				1.047	4.664**	4.362**
				(1.017)	(1.931)	(2.067)
Age in years	0.00618		0.00559	0.00697		0.00383
	(0.00437)		(0.00530)	(0.00434)		(0.00595)
Male	-0.239**			-0.255***		
	(0.0966)			(0.0933)		
Verbal IQ test	-0.0148		-0.0311	-0.00328		-0.0256
	(0.0305)		(0.0354)	(0.0322)		(0.0361)
Observations	66	50	50	66	50	50
Pseudo R-squared	0.117	0.067	0.099	0.108	0.115	0.132

Notes: Table presents results from a logistic regression of non-response on basic covariates and proportion of incorrect responses. Non-response=1 is defined as not responding in round 2 for the treatment group and not responding in round 3 for the phase-in group. Average proportion incorrect is the number of SMS sent back from the learner to the platform to complete exercises that are incorrect. Marginal effects are reported. Robust standard errors in parentheses. ***, **, * denote statistical significance at the 1, 5, 10 percent levels, respectively.

Table A9 : Self-esteem of non-respondents

	Non-respondents	**Respondents**
	Mean	**Mean**
	(s.d.)	**(s.d.)**
	(1)	(2)
Rosenberg Self-Esteem Score (4-item subset)		
Baseline	8.71	8.32
	(1.87)	(1.78)
Last three weekly survey contacts	8.68	NA
	(1.62)	
Average over all weekly survey contacts	8.66	8.94
	(1.53)	(1.88)
Differences from baseline values		
Difference Row 1 Row 2	0.03	
	(0.44)	
Difference Row 1 Row 3	0.06	
	(0.37)	
Number of Observations		
Baseline	21	44
Last three weekly survey contacts	50	
Average over all weekly survey contacts	155	636

Notes: The four panels presents mean values for the 4-item subset of the Rosenberg Self-esteem Scale for non-respondents (Column 1) and respondents (Column 2) who are receiving treatment (initial treatment group and phase-in controls). The rows of the first part present mean values at baseline, the last three measurement of self-esteem before dropout and over the course of the interaction with the platform. The second part presents differences from the baseline values for non-respondents (which are insignificant). ***, **, * would denote statistical significance at the 1, 5, 10 percent levels, respectively.

1 Methodological Appendix - Not for publication

This appendix outlines the methods that can be used to deal with attrition in a phase-in design. We first describe traditional approaches, followed by a series of non-parametric methods, including Lee bounds, monotone treatment selection and monotone treatment response. We then outline the assumptions under which Lee bounds can be tightened, and provide a proof. Finally, we show how to implement monotone treatment selection (MTS) and monotone treatment response (MTR) within a phase-in design.

1.1 Framework

Define the variables of interest $(Y_t^{1*}, Y_t^{0*}, S_t^{1*}, S_t^{0*}, T_t)$. The subscript is used to denote time, and the superscript treatment status. Y_t^{1*} and Y_t^{0*} are the outcomes of interest (reading skills and levels of empowerment) in the case where an observation is treated (Y_t^{1*}) or untreated (Y_t^{0*}). Since both states of the world are never simultaneously observed, Y_t^{1*}, Y_t^{0*} are latent or potential outcomes. S_t^{1*} and S_t^{0*} denote latent binary variables indicating whether an observation would be observed had it been assigned to treatment (S_t^{1*}) or not (S_t^{0*}).

The structure of the data is presented in Appendix Table A5 and Figure A1. There are three surveys, $t = 1$ is the baseline, $t = 2$ is the survey after the first phase of the program when the treatment group received treatment, $t = 3$ is the second phase of the treatment when the former control group (now cross-over group) receives treatment. T_t is an indicator for treatment assignment status for the period just prior to and ending in t. For example $T_2 = 1$ indicates the participant had received treatment just prior to survey round 2. The phase-in leads to the following structure of treatment assignment: $T_2 \in \{0, 1\}$, and $T_3 = 1$ if $T_2 = 0$ and $S_2^0 = 1$.[1] We do not observe the outcomes of the phase 1 treatment group in

[1] We do not observe any observations in phase 1 that were not treated. There are three observations in the control group, who did not participate in phase 2 of Cell-Ed, which have $S_2^0 = 1$.

$t = 3$. We note that Y_2^{1*}, S_2^{1*}, T_3 refer to latent variables of observations who are treated by survey round 3 but were in the original control group.

With regards to the outcomes, we observe

$$Y_2 = \begin{cases} Y_2^{1*} & \text{if } T_2 = 1, S_2^{1*} = 1 \\ Y_2^{0*} & \text{if } T_2 = 0, S_2^{0*} = 1 \\ missing & \text{otherwise} \end{cases}$$

$$Y_3 = \begin{cases} Y_3^{1*} & \text{if } T_3 = 1, S_3^{1*} = 1 \\ missing & \text{otherwise} \end{cases}$$

$$\begin{aligned} S_2 &= S_2^{1*} T_2 + S_2^{0*}(1 - T_2) \\ S_3 &= S_3^{1*} T_3 \end{aligned}$$

Given our randomization in treatment and control groups, $(Y_2^{1*}, Y_2^{0*}, S_2^{1*}, S_2^{0*})$, the outcomes (and observability) of interest, is by construction independent of T.

1.2 Parameter of interest

Our main parameter of interest is the *average treatment effect* (ATE) defined as: $E[Y^{1*} - Y^{0*}]$. In our application, where the minimum time period spanned by interactions with the platform was 41 days for phase 1 and 20 days for phase 2, all observations in the treatment group are (at least minimally) treated. Because of survey attrition, some observations on Y are missing. The parameter $E[Y_2^1 | T_2 = 1, S_2^1 = 1] - E[Y_2^1 | T_2 = 0, S_2^0 = 1]$ is the difference between the outcomes of observed treatment and control observations. This parameter, which we present in Table 3, does not identify the ATE unless the attrition is random, an assumption called the missing at random (MAR) assumption. If we are unwilling to make that assumption (or to assume that attrition is random *conditional on covariates*) then we need to take into account the selection into the survey.

1.3 Parametric selection models

The original parametric selection model (Heckman 1976) involves specifying the joint distribution of the errors in the selection and outcome equation as normal. The model is identified through this (very strong) functional form assumption. With an instrumental variable the identification of the parameter does not rely on the assumption of normality. In our context, there is no reasonable instrumental variable that meets the exclusion restriction, so we do not consider this approach.

1.4 Non-parametric approaches

Absent strong additional assumptions we cannot identify the ATE. With substantially weaker assumptions, we can, however, bound the parameters using non-parametric approaches.

1.4.1 Manski bounds

A first approach that does not rely on imposing assumptions on the selection process are so-called Manski bounds (Manski 1990; Horowitz and Manski 2000). Manski bounds are formed by imputing lowest and highest possible vales for the missing observations. For example, in our context, we could take the lowest and highest possible values on the test. To construct the lower Manski bound, we could impute the highest value possible for the control non-respondents, and the lowest possible value for the treatment non-respondents. For the upper Manski bound, we could impute the highest values for the treatment non-respondents and the lowest for the control group non-respondents.

We could tighten these bounds somewhat if we are willing to assume that the *observed* outcomes span the full support of outcomes. Then, the observed maximum and minimum values observed in round 2 (or across all of our surveys) can be taken as highest and lowest values. It is perhaps not unreasonable to assume that the top grade observed (or observed over three survey rounds) is the maximum a non-respondent might have achieved. As in most other applications, when we implement Manski bounds, the bounds are so wide as to be uninformative.

1.4.2 Lee (2009) bounds

Lee bounds make additional assumptions relative to Manski bounds. The first is the assumption of random assignment into treatment and control group, an assumption which is met in our context. Lee bounds are then based on assuming monotonicity in the selection into the survey: Either $S_t^{1*} \geq S_t^{0*}$ or $S_t^{1*} \leq S_t^{0*}$ with probability 1. This implies that treatment affects survey attrition in only one direction.

Consider the case $S_t^{1*} \leq S_t^{0*}$, that is, survey attrition is higher among treatment observations. This is the relevant case for our intervention. Define $y_q = G^{-1}(q)$ with G the cdf of Y, conditional on $T = 0, S_2^0 = 1$. y_q is the outcome of the qth quantile of the observed control group distribution. Further define

$$p_0 = \frac{P[S_2^0 = 1|T_2 = 0] - P[S_2^1 = 1|T_2 = 1]}{P[S_2^1|T_2 = 0]}$$

which is the additional survey attrition rate of the treatment group (relative to the proportion of the treatment group that is observed).

Lee (2009) shows that, under the monotonicity assumption, Δ_0^{LB} and Δ_0^{UB} are sharp bounds for the average treatment effect, where

$$\begin{aligned}\Delta_0^{LB} &= E[Y_2^1|T=1, S_2^1=1] - E[Y_2^0|T=0, S_2^0=1, Y \geq y_{p0}] \\ \Delta_0^{UB} &= E[Y_2^1|T=1, S_2^1=1] - E[Y_2^0|T=0, S_2^0=1, Y \leq y_{1-p0}]\end{aligned} \tag{1}$$

For example, if there is 15 percent higher survey attrition in the treatment group than in the control group, then the upper and lower bounds are created, by discarding, respectively, the 15% best and 15% worst outcomes in the observed control. Imposing monotonicity ensures that no observed treatment observation would have been a missing observation if they had not received treatment (because treatment only affects attrition in one direction - here it increases it). Thus, all the individuals who make up the 15 percent extra non-respondents from the treatment group would have been observed had they not been treated. But we don't know which ones they are, hence we trim the best and worst observations.

Tighter bounds

Lee (2009) proves that bounds conditioning on covariates are (weakly) tighter. He notes that bounds do not tighten if the trimming proportion in each subgroup is exactly the same, as then the same observations are trimmed regardless of conditioning on the covariate.[2]

We show that, under an additional assumption, conditioning on the phase-in information will, except for an implausible case, always narrow bounds. The additional assumption is that $P(S_2^{1*} = 1|S_3^1 = 1) = 1$ or $P(S_3^{0*} = 0|S_2^1 = 0) = 1$. These two formulations are equivalent.The first part is the assumption that those who participated in the survey later in round 3 would have participated in the survey in round 2 if they had received treatment

[2] This is not the only case, however. Define G as the cdf of the group with less attrition, N_s as the excess attrition trimmed, with N the corresponding number trimmed. N_{as}, N_a denote the relevant subgroup equivalent, created by the covariate (say gender). Then bounds always tighten unless $G_a(G^{-1}(\frac{N_s}{N})) = \frac{N_{as}}{N_a}$, i.e. the excess attrition in every group taken by itself is equivalent to the trimming proportion in that group without conditioning. More intuitively, conditioning on covariates tightens bounds, when the group by group excess attrition is not perfectly correlated with the share of that group in the top and bottom groups.

immediately. This is equivalent to the second part which specifies that that those who do not respond to the survey after treatment in round 2 would also not have participated if they had first been in the control group and then received the treatment later.[3] This assumption cannot be tested, and may, for example, be violated when there are time-variant determinants of survey participation. [4]

The assumption allows us to use the phase-in information to identify the group of observations in the control group that the treatment non-respondents would be part of (or in the weaker version would be equivalent to). Lee's trimming methodology implies trimming observations in that subset of the control group. The reason the phase-in always tightens bounds is that overall trimming is now completely concentrated in one of discrete groups (in our case one of two groups formed by survey response and non-response). We show that this always tightens the bounds in the following proposition:

Proposition 1 *Take a population g of size N, split into two groups g_i of sizes $N_i, i = 1, 2$. Without loss of generality, let $N_s < min(N_1, N_2)$ and define $max_{N_s(h)}$ as the N_s largest observations in a group h. Define $min_{N_s}(h)$ similarly. Then for any $i \in 1, 2$*

1. $\frac{1}{N_s} \sum min_{N_s}(g_i) \geq \frac{1}{N_s} \sum min(g_a \cup g_b)$*, where the summation is taken over the observations in the set. The inequality is strict, unless* $G_b(G_a^{-1}(\frac{N_s}{N_a})) = 0$*. This is only the case if the N_s smallest observations in group i lie below the minimum observation in group b.*

2. $\frac{1}{N_s} \sum max_{N_s}(g_i) \leq \frac{1}{N_s} \sum max(g_a \cup g_b)$*, where the summation is again taken over the observations in the set. The inequality is strict, unless* $G_b(G_a^{-1}(\frac{1-N_s}{N_a})) = 0$*. This is only the case if the N_s largest observations in group a lie above the maximum observation in group b.*

[3] If there was no control group non-response in the first round, this would simplify to $S_2^{1*} = S_3^{1*}$

[4] A stronger assumption would be that, in expectation, the latent round 2 outcomes of those who would have participated in survey round 2, but not in round 3 are the same as of those who would have participated in round 3 but not in round 2: $E[Y^{0*}|S_2 = 1, S_3 = 0] = E[Y^{0*}|S_2 = 0, S_3 = 1]$.

The exceptions to the tightening of the bounds are theoretically possible, even if extremely implausible for any real-world situation. For both bounds not to be changed, there must be no overlap in the distribution of the observed later participants and non-respondents in the control group (in the percentiles that are trimmed). This is, unsurprisingly, rejected by our data.

The approach to dealing with attrition extends to situations where there is survey *and* program attrition. In that case the information from the control group would be separated by whether the person is just a survey non-respondent or also does not take up the program in addition to not responding on the survey.

The methodology is even more flexible, as it allows for conditioning on a wide range of post-intervention covariates, if the above assumptions are plausible. In our particular application, we have information on platform usage during the treatment. We can condition on the amount of time learners spent interacting with Cell-Ed, or the highest module completed. In our application, we are unable to implement this idea as the sample size is too small.[5]

1.4.3 Monotone treatment response, monotone treatment selection and monotone IV

In this sub-section we turn to alternative assumptions that also lead to a tightening of the bounds around the treatment effect.

Monotone treatment selection

The monotone treatment selection assumption (Manski and Pepper, 2000) places structure on the selection mechanism, namely that survey non-respondents have lower outcomes on average than those who take up treatment. It assumes that those who select into the treatment have higher expected outcomes under both treatment and non-treatment than those

[5] Note that DiNardo et al. (2006) also make use of post-treatment data, but of a different sort: they show how to create and use variation in survey response rates to deal with attrition. Instead we identify who are the equivalent non-respondents in the control group.

who do not. In our context, we apply this assumption to mean that those who do not even respond to the survey in round 2 have lower outcomes than those who *stick around* knowing that they will receive treatment after round 2. This can be summarized as:

$$E[Y_2^0|S^{0*} = 0] \quad < \quad E[Y_2^0|S^{0*} = 1] \tag{2}$$

This allows us to bound from above the outcomes of the three control group always non-respondents during round 2, relative to control group cross-over non-respondents.[6]

Monotone treatment response

The monotone treatment response assumption specifies that treatment cannot make anyone worse off, or $Y^1 > Y^0$. We implement monotone treatment response by bounding missing treatment observations' round 2 outcome with their baseline outcome, that is $Y_2^1 > Y_1^0$. We note that while $Y^1 > Y^0$ cannot be tested, the assumption $Y_2^1 > Y_1^0$ can.[7'8]

We implement the MTR and MTS assumptions jointly to estimate lower bounds on the treatment effects in the regressions contained in Table 5.[9]

[6] With treatment effect heterogeneity, a natural alternative interpretation of marginal treatment selection is that the non-respondents are those with lower treatment effects.

[7] For observations in the treatment group that we observe in both round 1 and round 2, we only reject that $Y_2^1|S_2^1 = 1 > Y_1^0|S_2^1 = 1$ for one single observation. In that particular case, we observe that after treatment, the test score decreases by 0.13 for basic reading (compared with the lowest lower bound of the effect 1.068). There is also a very high correlation among literacy scores between round 1 and round 2 for the control group, of around 0.9, with almost no difference in means

[8] A weaker version, called the Mean Monotone Treatment Response (MMTR), implies that observations in the treatment group on average have at least as good outcomes as if they had not been treated, and this is true for subgroups of observations as well. For our context it means that the average outcome for non-respondents in the treatment group is larger than for the observations in the control group that will be non-respondents in round 2.

[9] For the control observations, we make a slightly more conservative assumption than just MTS. We bound the observations from above by the maximum of the following: the minimum of the cross-over non-respondents round 2 outcomes and their own baseline outcome.