Toward Learning at Scale in Developing Countries: Lessons from the Global Learning XPRIZE Field Study

Andrew A. McReynolds

U. of San Francisco
San Francisco, CA, U.S.
AAMcReynolds@usfca.edu

Sheba P. Naderzad

U. of Southern California
Los Angeles, CA, U.S.
naderzad@usc.edu

Mononito Goswami

Delhi Technological U.
New Delhi, India
mononito_bt2k16@dtu.ac.in

Jack Mostow

Carnegie Mellon U.
Pittsburgh, PA, U.S.
mostow@cs.cmu.edu

ABSTRACT

Advances in education technology are enabling tremendous advances in learning at scale. However, they typically assume resources taken for granted in developed countries, including reliable electricity, high-bandwidth Internet access, fast WiFi, powerful computers, sophisticated sensors, and expert technical support to keep it all working. This paper examines these assumptions in the context of a massive test of learning at scale in a developing country. We examine each assumption, how it was broken, and some workarounds used in a 15-month-long independent controlled evaluation of pre- to posttest learning and social-emotional gains by over 2,000 children in 168 villages in Tanzania. We analyze those gains to characterize who gained how much, using test score data, social-emotional measures, and detailed logs from RoboTutor. We quantify the relative impact of pretest scores, literate aspirations, treatment, and usage on learning gains.

Author Keywords

Global Learning XPRIZE; Tablet tutors; Developing countries; Literacy; Numeracy; Social-emotional; UNESCO.

CSS Concepts

•Applied computing~Education~Interactive learning environments •Applied computing~Education~Computer-assisted instruction •Applied computing~Education~Distance learning

INTRODUCTION

The Global Learning XPRIZE (learning.xprize.org) was an unprecedented experiment in learning at scale – a four-year, \$15M worldwide competition to develop an open-source Android tablet app to teach basic reading, writing, and numeracy to children with little or no access to schools. It was a pilot proof of concept to test whether tablet apps could overcome the failure of market forces to provide basic education to the world's 250 million children who have no school, and the much larger number deprived of effective instruction.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

L@S '20, August 12–14, 2020, Virtual Event, USA.

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7951-9/20/08...\$15.00.

DOI: <https://doi.org/10.1145/3386527.3405920>

In 2015, 198 teams from 40 countries registered to compete in the Global Learning XPRIZE. In 2017, 38 of them submitted their apps, in both English and Swahili as required. XPRIZE installed the apps on Google Pixel C tablets and sent them to its panel of independent judges. The judges ran the apps, reviewed teams' supplemental materials, and selected 11 semifinalists. XPRIZE subjected each semifinalist's Swahili app to an hour of video-recorded individual play testing by 15 children at a school in Dar es Salaam.

Based on the recorded play tests, semifinalists' presentations to the panel, and their answers to its questions, the judges selected five teams as Finalists. Each Finalist won \$1 million USD and competed for the \$10 million Grand Prize to be awarded to whichever team(s) not only achieved the highest total learning gains in a 15-month-long independent controlled evaluation study but significantly outgained a control group in reading, writing, and numeracy, weighted respectively as 60%, 10%, and 30% of the total score.

In contrast to previous XPRIZE competitions, the Global Learning XPRIZE was the first one to address education, take place in Africa, focus on software, or require open source.

The rest of this paper is organized as follows.

First, we identify six assumptions common in learning at scale, how they are violated in developing countries, and how XPRIZE and the Finalists addressed each one. We relate the XPRIZE study to previous studies.

Study 1 describes the overall results of the Global Learning XPRIZE Competition, and analyzes them to characterize who benefited how much.

Study 2 dives deeper by focusing on one Finalist (RoboTutor) and extending our analysis to incorporate its detailed log data.

Finally, we conclude by summarizing contributions, limitations, and future work.

BROKEN ASSUMPTIONS

Typically, modern educational software is developed in, by, and for WEIRD (Western, Educated, Industrialized, Rich, Democratic) countries, and therefore assumes reliable electricity, high-bandwidth Internet access, fast WiFi, powerful computers, sophisticated sensors, and expert

technical support. Research to evaluate this software relies even more heavily on such assumptions due to the additional instrumentation and data collection involved. These assumptions are often violated in developing countries. We now discuss each assumption, how it is violated, and how XPRIZE and the Finalists' apps addressed it, whether by enabling it or by working around it.

Electrical Power

Educational apps typically assume reliable electricity. In developing countries, the reality tends to be quite different, with limited availability and fluctuating voltage.

The obvious and widespread workaround is to use batteries, but batteries must be replaced or recharged. XPRIZE worked around this limitation by providing a solar-powered recharging station for each village. Another solution is to include a small solar recharger in or with each tablet.

An unexpected problem with the Pixel C tablet battery arose whenever it ran out of power. Recharging the battery reset the tablet clock to 12/31/1999, rendering logged timestamps invalid and wreaking havoc on analysis of log data. One team's workaround to this problem was to advance the clock by a very large time interval (e.g. a year) after each clock reset and increase it by a large amount each time. Our workaround was to use the week that XPRIZE staff uploaded the log files to deduce lower and upper bounds on when they were actually created.

Internet

Many educational apps assume reliable, wide-coverage, high-bandwidth Internet access. The reality in developing countries is that Internet access is limited, slow, or non-existent.

The obvious and necessary workaround is apps that run without relying on an Internet connection. Remote villages in the XPRIZE study had no Internet connection. Cell phone service was sometimes available, but with bandwidth too low to transmit large quantities of data. XPRIZE's workaround was for its field staff to collect log data on portable drives and send it back via motorcycle to a location with a good Internet connection. From there, XPRIZE staff uploaded each team's data to the Box cloud server for the team to download.

WiFi

Similarly, many apps assume reliable, wide-coverage, high-bandwidth WiFi. The reality in developing countries is that WiFi, when available at all, has frequent outages, limited coverage, and low bandwidth.

XPRIZE's workaround was to provide a solar-powered server for each village so that tablets could transfer files to and from it while within WiFi range, typically while recharging. XPRIZE field staff collected files from village servers on their weekly visits to each village.

The XPRIZE apps' workaround was to run client-side without relying on the local server, but take advantage of WiFi when available to transfer data, in particular log files.

However, fluctuating WiFi access apparently prevented transfer of an unknown number of files. Our partial workaround for unknown data loss was for RoboTutor to number log files sequentially and include their sequence numbers in their filenames. This remedy did not eliminate the data loss, but at least the gaps in the log numbering sequence exposed the number of missing logs. Modeling the missing logs as a random sample of the complete set made it possible to work around data loss by extrapolating analyses based on the sample.

The form and feasibility of such extrapolation varied. For example, the session duration statistics computed from the sample applied to the original set as well. In contrast, extrapolating usage frequency statistics required adjusting them to account for the amount of missing data. If 10% of logs were lost, then session frequency based on the rest of them should be multiplied by 10/9. However, longitudinal analyses are problematic to extrapolate. For instance, knowledge tracing models are based on observing a complete series of opportunities to use a skill. The impact of missing observations depends on their position in the series.

Computers

Educational software in WEIRD countries can assume that computers are plentiful, powerful, and personal. XPRIZE enabled this assumption by providing a high-end tablet for every child at its study sites and replacing it if lost, stolen, or broken. Google donated 8,000 Pixel C tablets for the study.

The reality is that computers in developing countries are typically scarce, low-end, and shared. Tablet sharing is not an issue for non-adaptive apps, but it is a problem for apps that depend on tracking individual progress to adapt to the student.

The RoboTutor team user-tested successive versions at several beta sites in Tanzania but could not afford to provide a separate tablet for each child. As a workaround, it supported tablet sharing by developing a self-enrollment and login mechanism called FaceLogin (shown in Figure 1 below) that children could operate without being able to read, write, or get help from anyone who could.

Figure 1: Screenshot of FaceLogin on a shared tablet

FaceLogin prompted a new user to say his or her name aloud, recorded a 3-second video of it, and added the child's photo to an on-screen gallery of enrolled users. Thereafter the child could log in simply by tapping on the photo to see and hear the video and tapping again to confirm. Field tests of this mechanism succeeded a bit too well at first: children loved taking selfies so much that they kept re-enrolling, thereby defeating its purpose. To remove the incentive to re-enroll, the team modified the mechanism to let already-enrolled children take recreational selfies without re-enrolling. To work around the poor quality of some selfie photos evident in Figure 1, for example due to backlighting, FaceLogin also prompted children to choose animal photos to display next to them to make them visually distinct. FaceLogin should therefore be modified to keep multiple children from picking the same animal, such as the apparently overly popular bat.

An alternative workaround developed by another team used continual assessment in its app to adapt quickly when the tablet changed hands. This approach eliminated enrollment but required adaptation based on scant data from the student.

Sensors

Advanced educational software typically assumes that sensors are reliable, accurate, and many-modal, i.e. exist for diverse forms of input (audio, vision, touch), and are used in a controlled environment.

The reality in developing countries is that sensors are typically unreliable, inaccurate, and few in number, and used in noisy environments.

Various workarounds can help, e.g.:

- Apply the sensor selectively, i.e. only in situations where it can work. For instance, RoboTutor uses speech recognition only for highly predictable speech, such as listening to oral reading of a known text or listening for an expected answer to a problem.
- Degrade gracefully. For instance, read a word aloud to the child after two attempts by the child to read it.
- Provide a fallback. For instance, let the child tap to make RoboTutor read the next word.

Support

Educational software is typically used in environments where expert assistance and assessment – both technical and educational – are available when needed. For example, a teacher or parent can help a student who has difficulty understanding the content or operating the software and assess whether it is at an appropriate level for the student.

In developing countries, such expertise is often scarce or absent. Available adults are inexperienced or even illiterate. Teachers are over-burdened, under-trained, absent, or non-existent.

Workarounds used by Finalists' apps included verbally prompting the student what to do and pointing with an animated hand to indicate a target to look at or tap, because

just highlighting it is not self-explanatory to users unfamiliar with touchscreens – the sort of cultural difference revealed by in-person study of the context of use. Such animations are useful within the context of an app, but video provides a useful workaround for unavailable human assistance by displaying external context. For instance, RoboTutor plays short video tutorials of humans demonstrating how to hold the tablet and operate various activities in order to work around the lack of someone to do it in person.

In the absence of a human teacher to assess student knowledge, Finalists' apps used various workarounds. One simplistic workaround is to progress through the curriculum in a fixed order starting from the easiest lesson. This workaround relies on the unrealistic simplifying assumption that the resulting sequence will fit every student's rate of learning. Another workaround is to let students choose what to do. This "learner control" makes the unrealistic assumption that children's goal is to optimize their learning, and that they know which choice will do so. This assumption can be mitigated somewhat by restricting the choices to activities whose prerequisites the student has completed.

A more sophisticated workaround is to automate assessment. Automated placement determines where in the curriculum the student should start. Automated promotion determines when the student has made sufficient progress to advance. Automated assessment uses and updates a student model.

Relation to Prior Work

We now relate the Global Learning XPRIZE to previous large-scale, rigorously controlled evaluations of education technology for young children. Such studies are rare.

The relatively few such studies that have been done were in WEIRD settings, notably the effectiveness studies of children's literacy and math interventions reviewed by the United States Department of Education's *What Works Clearinghouse* (ies.ed.gov/ncee/wwc). Only a subset of these studies evaluated education technology for young children. They generally ranged from dozens to hundreds of participants, compared to over 2,000 in the Global Learning XPRIZE.

Some controlled evaluations of children's education technology have been carried out in non-WEIRD settings, for example in Ghana [5], India [2, 11], and Tanzania [8, 9]. These studies involved up to a few hundred children, and generally lasted at most a few weeks or months.

Some experiments have studied MOOCs in non-WEIRD settings [4, 6]. However, MOOCs require reliable Internet access, which has extremely limited availability in developing countries. Moreover, MOOCs are for older students, not young children.

Duolingo is used around the world, but teaches languages, not basic numeracy or (until very recently) literacy. Although it is used by millions of language learners and conducts A/B testing to evaluate new features, a third-party evaluation of its

effectiveness [10] was conducted with a much smaller sample of several hundred English-speaking Spanish learners recruited online from the United States. The study measured pre-post gains but had no control group.

In short, the Global Learning XPRIZE study was unusual in its setting (remote villages in Tanzania), size (thousands of participants), target users (out-of-school children 9-11 years old), duration (15 months), design (five treatment groups and a control group), and platform (Android tablets).

STUDY 1: OVERALL ANALYSIS OF GAINS

XPRIZE hired [DataVision International](#) to pre- and posttest study participants, and [RTI International](#) to analyze the results summarized in this section and in RTI's 5/16/2019 presentation [3]. These results are based on a total of 2041 participants from 168 villages in the Tanga region of Tanzania, remote from a school or city center, including 30 villages assigned to a control group. 74% of the participants had never attended school, and 80% were never read to at home. They were mostly 9-11 years old, with 8% under 9 years and 3% over 11 years.

XPRIZE selected six groups of villages to be demographically similar and assigned each of the five Finalist apps to be used in a different group of villages. The other 30 villages served as a delayed-treatment control group that was also pre- and posttested but received the same “business as usual” instruction (or lack thereof) as before the study, with no tablets. After the study, the control villages were to receive the winning app.

Measures

Testers from DataVision International administered the Elementary Grade Reading Assessment (EGRA, at shared.rti.org/content/early-grade-reading-assessment-egra-toolkit-second-edition) and the Elementary Grade Mathematics Assessment (EGMA, at shared.rti.org/content/early-grade-mathematics-assessment-egma-toolkit) in Swahili to children one-on-one before and after the 15-month treatment period, without knowing which, if any, app they were assigned to use.

Testers also recorded demographic information such as age, gender, and prior schooling. For privacy, XPRIZE used anonymized identifiers for villages and participants.

[RTI International](#) (blind to which data came from which group) calculated scores in reading, writing, and numeracy as the percentage of correct answers on their respective subtests, and gains as posttest minus pretest [3].

In addition, 38 research assistants, mostly graduate students from the University of Dar es Salaam, administered a Social-Emotional Assessment at baseline, midline, and endline for a UNESCO (United Nations Educational, Scientific and Cultural Organization) project. Its goal was to “understand the impact of the project on out of school children’s social emotional development” of “aspiration, sharing/conflict solving, honesty, persistence, self-regulation, and

obedience/respect” [7]. UNESCO and its partners at the University of Dar es Salaam formulated what they thought were the most important questions for their community. The Social-Emotional Assessment used a structured face-to-face questionnaire with the child and their parents [7].

The questionnaire elicited children’s future career aspirations by asking (in Swahili), “Please tell me when you grow up, who would you like to become?” Table 1 shows the distribution of the 1183 responses from the baseline and endline assessments. Borrowing a scheme from XPRIZE for coding aspirations, we classified them subjectively and somewhat arbitrarily as either literate (e.g. teacher or doctor) or non-literate (e.g. farmer or housekeeper) based on whether they require literacy.

Literate Aspirations	N	Non-Literate Aspirations	N
Teacher	454	Farmer	355
Businessperson	161	Not Sure	343
Driver	155	House Keeping	178
Doctor	131	Footballer	92
Police	60	Livestock Keeping	64
Nurse	41	Employee	52
Technician	27	Soldier	51
Studying	18	Mechanic	31
Politician	15	Cook	17
Religious Leader	13	Athlete	13
Engineer	10	Mason	12
Pilot	8	Tailor	9
Scientist	3	Parent	8
Journalist	3	Go to Town	6
Musician	2	Guard	6
Researcher	2	Motorcycle Rider	6
English Teacher	1	Hotelia	5
President	1	Village Leader	4
		Artist	2
		Getting Married	2
		Pastoralist	2
		Brother	1
		Veterinary Officer	1

Table 1: Children’s future aspirations

The questionnaire assessed positive behavioral attributes (e.g. sharing) by asking, e.g., “Imagine a situation where you are playing with a toy or a ball or anything that you like and another child wants to play with it. What would you do in this situation?” A pre-generated list unknown to the child classified anticipated positive responses. For example, it

classified “take turns” or “play together” as sharing and “get another ball,” “run away with the toy,” “hit him/her,” “tell the child it’s mine and he cannot have it,” or any other negative response as not sharing. We scored responses as 1 (positive) or 0 (negative). We will refer to these ratings, e.g. “persistent” and “cooperative,” as if they accurately captured the constructs they were intended to measure.

XPRIIZE Data Set

XPRIIZE provided test scores for the five pooled treatment groups with village and participant identifiers removed. It also provided item-level test performance and responses to the social-emotional questionnaires for 75% of the study participants assigned to each group, including their anonymized identifiers. The data is publicly available at data.xprize.org.

XPRIIZE Results

Over the course of the 15-month controlled study, the control group ($n=361$) score rose from mean 8% (sd 9%) at pretest to mean 16% (sd 15%) at posttest, gaining mean 8% (sd 11%). In contrast, gains for the 5 Finalist teams averaged 2-3x higher, ranging from 18% (sd 16%) to 26% (sd 17%). Thus, the five Finalists’ apps beat the control group not just significantly ($p < .000$ on unpaired T tests, $n = 2041$) but dramatically (effect size .82), demonstrating that tablet apps have great potential to provide basic education to children who do not currently receive it.

Individual Differences

Although the Finalists far surpassed the control in mean test score gains, not all participants benefited equally. Far from it! A substantial percentage had negligible, zero, or even negative gains from pre- to posttest, as shown by the large number of points near, on, or below the horizontal axis in Figure 2, which plots the gains of the 1680 children in the five treatment (non-control) groups on the vertical axis against their pretest scores on the horizontal axis. Individual gains ranged from -32% to $+52\%$ in the control group and from -11% to $+91\%$ in the groups using the Finalists’ apps. Thus, although impressively successful overall, the apps left considerable room for improvement.

It is therefore important to analyze which children gained much more or less than others, and why. Accordingly, we analyzed the results in an attempt to characterize who did how well, as we now describe.

Analysis of Variability

To shed light on the distribution of gains, we built statistical models to predict them from the available predictors. Condition (Treatment or Control) was a predictor. So were pretest scores and the social-emotional baseline measures. Other predictors such as gender, age, and prior schooling came from demographic data. Finally, as a proxy for village size, we used the number of study participants in the village as an aspect of village social structure that might affect children’s learning.

Additional predictors are possible. For instance, children’s self-reported chores (e.g. tending goats) might shed light on their amount of usage, but using them as predictors would require quantifying them, e.g. as the number of chores or an estimated number of hours required.

We included all the predictors in multiple linear regression models to predict gains in test scores and social-emotional measures. We then used backward regression to eliminate the non-significant predictors from each model. Each column of Table 2 summarizes the resulting model of a gain for which Condition was a significant predictor. The table has a row for each baseline predictor that was significant in at least one model. Gender, age, village size, prior schooling, sharing, honesty, and self-regulation were not significant predictors of gains in any of the measures.

Figure 2: Scatterplot of Score Gains on Y-axis vs. Pretest Scores (% correct) on X-axis, with treatment group ($N = 1680$) in red and control group ($N = 361$) in blue.

Baseline Measures	Gain Measures			
	Test Score	Persistence		Honesty
	$\bar{x}$: 20.27	31% ↑††	69% ↓†††	74% ↑†† 21% ↓†††
Condition	13.39**	0.09*		0.10**
Test Score	0.14**	-0.004**		
Aspiration ($\bar{x}$: 0.42) [†]	4.34***			
Persistence ($\bar{x}$: 0.26) [†]		-1.00***		0.08**
Honesty ($\bar{x}$: 0.62) [†]				-0.96**
Intercept	5.88	0.35		0.64
Adjusted R ²	0.11	0.48		0.56

[†] Mean social-emotional measure score at baseline.

*p < .05; **p < .01; ***p < .001

†† Percentage of children whose score on this social-emotional measure rose from 0 at baseline to 1 at endline

††† Percentage of children whose score on this social-emotional measure fell from 1 at baseline to 0 at endline

Table 2: Summary of regression models of individual differences in gains

For example, three baseline measures were significant predictors of test score gains, leading to this model:

$$\begin{aligned} \text{Score Gain} = & 5.88 + \\ & 13.39 * \text{Condition} + \\ & 0.14 * \text{Pretest} + \\ & 4.34 * \text{Aspiration} \end{aligned}$$

This regression equation is significant ($F(3, 1106) = 43.44, p < .000$), Adjusted $R^2 = .105$. Here, *Score Gain* is *Posttest* minus *Pretest*, both measured as % correct; *Condition* is 1 for Treatment (i.e. using one of the Finalists' apps) and 0 for Control; and *Aspiration* is 1 for literate and 0 for non-literate. So, on average:

- Treatment increased test score gain by 13.39 points.
- Each additional point (% correct) on Pretest was worth an additional gain of 0.14 points on Posttest.
- Children with literate aspirations at baseline gained 4.34 points more than children with non-literate aspirations.

We were interested primarily in predicting gains in test scores, but also identified predictors of social-emotional gains. Most such predictors were very weak, representing statistically reliable but minor effects. In particular, *Condition* had marginal effect on social-emotional measures; the XPRIZE apps had little of the hoped-for impact on those measures, unlike their dramatic effect on learning gains.

Baseline values were strong negative predictors of gains on two of the social-emotional measures, but they may simply reflect the fact that gains for children who started with positive values were necessarily negative because they had nowhere left to go but down. In particular, the -1.00 regression coefficient for the relation of baseline to gains for persistence reflects the fact that of the 277 children rated

as persistent at baseline, 190 (69%) were rated non-persistent at endline and 87 as still persistent. Of the 800 children rated as non-persistent at baseline, only 248 (31%) were rated persistent at endline and 552 as still non-persistent. Likewise, the -0.96 coefficient for honesty gains vs. baseline reflects the fact that 141 (21%) of the 666 children rated as honest at baseline were rated as not honest at endline. 303 (74%) of the 409 children rated as not honest at baseline were rated as honest at endline. Although there were more positive than negative gains in both persistence and honesty, the imbalance would have had to be greater to outweigh their respective intercepts of 0.35 and 0.64 for the coefficients to be positive.

STUDY 2: ANALYSIS OF ROBOTUTOR USAGE DATA

Next, we focused on RoboTutor (shown in Figure 3) and augmented our set of predictors by adding some metrics quantifying individual children's usage of its final, "code drop 2" version, deployed for the last three months of the 15-month study. Some children had stopped using RoboTutor by this point, so we omitted them from the analysis of usage.

Figure 3: Screenshot from RoboTutor numeracy activity

Apparatus

This section focuses on analysis of results, so a more detailed description of RoboTutor is beyond its scope; for more information, see www.robotutor.org. Figure 3 shows a screenshot of a bubble-popping activity for numeracy, in which RoboTutor spoke a number, e.g. 5, and the child's task was to tap on the bubble containing five dots. Launching RoboTutor started a session that lasted until the child exited or RoboTutor crashed. A session consisted of activities jointly chosen by RoboTutor and the child. The code drop 2 version of RoboTutor cycled among the three content areas (literacy, numeracy, literacy, stories, repeat...), offered two activities in the chosen area, and the child chose one. An activity (game or tutor) ended when the student completed it or tapped the *Back* button at the top left of the screen.

Variables

We analyzed logs to measure children's usage of educational activities (games or tutors) for teaching a given skill.

Duration would be the most obvious measure, but we did not use it, due to problems in calculating it. At one extreme, duration was zero for activities with only one attempt, because we computed the duration of an activity as the interval between the first and last attempts. At the opposite extreme were spurious multi-day durations caused by prolonged periods of user inactivity. Nevertheless, we get some idea of temporal grain size from the 24-second median attempt duration of activities with more than one attempt, with the caveat that attempts were much shorter for fast-paced activities like bubble-popping and longer for others.

Accordingly, we used three non-temporal measures of usage:

- The number of items completed
- The number of attempts to perform single-step items or steps in multi-step items.
- The number of correct attempts

To analyze the effect of activity usage on test scores, we were limited to the 166 children still using the app after a full year, only 138 of whom had baseline social-emotional scores.

Accordingly, we compared three models:

1. Predict all 166 children's subtest score gains from their activity usage without social-emotional scores.
2. Predict the 138 children's gains from their activity usage and social-emotional baseline scores.
3. Predict the 138 children's gains from their usage alone.

Comparing models 2 and 3 showed that social-emotional baseline variables made no difference except to predict that obedience added 16.73% to gains in quantifying. We concluded that social-emotional variables generally did not affect children's gains directly, only indirectly by affecting their usage. Accordingly, we used model 1 to predict gains from usage, to take advantage of the larger sample ($n=166$).

RoboTutor Usage Results

Table 3 shows the results of separate linear models to predict the gain on the EGRA or EGMA subtest of each skill from usage measures of the activities to teach that skill, pretest

Predictors	Test Score Gains							
	<i>Familiar Word</i>	<i>Invented Word</i>	<i>Dictation</i>	<i>Number Identification</i>	<i>Quantifying</i>	<i>Number Patterns</i>	<i>Word Problem</i>	<i>Level 1 Addition</i> <i>Level 2 Subtraction</i>
<i>Usage</i>								
# Items		0.16*		0.012*	0.039*	0.18*	17.20**	
# Attempts	-0.084**					-0.006*	90.97*	-1.34***
# Correct Attempts	0.11***		0.008*				-91.10*	0.01* 3.12***
<i>Baseline</i>								
Pretest Score	1.08***	0.50***	-0.51*					0.463**
Age			-4.70*					-3.28*
Village Size							-1.15*	
Prior Schooling			12.47*					
<i>Intercept</i>	4.90	4.55	79.62	20.94	13.66	8.97	26.73	44.04 2.37
<i>Adjusted R²</i>	0.17	0.08	0.09	0.03	0.02	0.03	0.07	0.04 0.18

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 3: Relation of gains on subtests of individual skills to usage of RoboTutor activities intended to teach them

score on the subtest, and demographic information. To attribute subtest gains to activities, we performed a cognitive task analysis to identify the skill(s) measured by each subtest and exercised by each activity.

Table 3 is challenging to interpret. All five coefficients on # of items are positive, with values ranging from gain per item completed of .012% for Number Identification up to 17.2% for Word Problems. However, there are negative coefficients on # attempts for Familiar Word, Number Patterns, and Level 2 Subtraction. Two of these negative coefficients are offset by positive coefficients on # correct attempts, suggesting that # incorrect attempts (not included in the model) is the true negative predictor. It is not clear how to explain the lack of such an offset for Number Patterns, but the negative coefficient is small (-0.006) and only weakly supported ($p < .05$). All but one coefficient on # correct attempts is positive except for the -91.10 on Word Problem, which is offset almost exactly by the 90.97 on # attempts.

To analyze effects of social-emotional variables on overall usage, we started with all of them as predictors, along with pretest and demographic variables as before, with # items, # attempts, and # correct attempts as outcome variables, and performed backwards regression by eliminating non-significant predictors.

Unfortunately, social-emotional variables were disappointing predictors of total usage. After dropping non-significant predictors, only one outcome variable remained (# items), with only two significant ($p < 0.05$) predictors, namely self-regulation and age. Surprisingly, the resulting model predicted that students would complete 697 *fewer* items if they were self-regulated at baseline, and that older students would complete 419 fewer items per year of age. Again, these variables might do better at predicting total usage over the entire 15 months of the study, rather than just the final three months.

CONCLUSION

We now summarize our contributions and future work.

Contributions

The Global Learning XPRIZE was a landmark experiment in education technology. We discussed how it addressed some assumptions violated in developing countries. As a proof of concept, it showed that educational tablet technology can provide basic education dramatically more effective than the little or none that the children would have received otherwise. Thus, it has great promise.

As we showed, however, gains were distributed very unevenly, with some children gaining greatly and others only slightly or not at all. Thus, the promise was realized only partially. Understanding why so many children learned little or nothing is essential to realizing the promise more fully. This paper contributes to that understanding by using linear regression models of test scores, social-emotional measures, and demographic variables to characterize who gained how much from pretest to posttest. We used linear regression

because of its interpretability compared to more powerful black box machine learning models such as Random Forests whose higher accuracy sacrifices understandability. Our goal was not to maximize accuracy, just to identify factors that affect gains and quantify their impact.

Aggregating across the five Finalists' apps showed that:

- Treatment boosted test score gains by over 13 points but had only marginal impact on social-emotional measures.
- Each additional pretest point added 0.14 points of gain.
- Children with literate aspirations at baseline gained 4.34 points more than children with non-literate aspirations.
- Gender did not predict gains in test scores or in social-emotional measures. On average, boys and girls gained equally, which is good news for promoting gender equity.
- Gains outnumbered losses in persistence and honesty.

Using RoboTutor log data to analyze usage of activities targeting specific EGRA and EGMA subtests showed that:

- Each additional subtest point increased subtest gains by 1.05 points for familiar words, 0.50 points for invented words, and 0.46 points for 2-digit subtraction.
- The number of attempts was a finer-grained and more reliable predictor of gains than the number of items.
- For some subtests, gains rose with the number of correct attempts but fell with the total number of attempts.

Limitations

An important caveat to keep in mind is that the test score gains were over the 15-month interval from pre- to posttest, but the usage measures analyzed here were only over the final three months. We would expect usage measures for the entire 15 months to be better predictors of test score gains.

We listed the variables considered, eliminated non-significant predictors, and reported the resulting models, but did not control for false discovery rate or multiple comparisons. Moreover, we explained only 2% to 18% of the considerable variance in children's gains, depending on subtest.

Future Work

This work could be extended in multiple directions, for example:

Test other models, such as with interactions. However, complex models can overfit, and interpreting interactions is especially hard for numerical variables such as age and score.

Operationalize usage quality metrics to measure effects on and of student engagement, persistence, and performance.

Perform finer-grained analyses of gains. We measured usage simply by counting students' items and attempts across all the activities relevant to each EGRA and EGMA subtest. In contrast, more precise analyses would trace learning curves for individual skills and items.

Analyze the relationship among context, usage, and gains.

Only two of the baseline social-emotional measures collected for the UNESCO study were significant predictors of gains or usage: literate aspirations predicted higher gains, and self-regulation (surprisingly!) predicted completion of *fewer* total items. Future field studies of tablet tutors and their component activities should examine their context in richer detail, including its effects on different measures of usage – duration (amount of usage per day used), frequency (days used per week), and longevity (total number of weeks used). That is, researchers should ask both adult observers and the children themselves what led to starting or stopping usage of the tutor or a particular activity at a given moment, returning to it on a later day, and continuing or discontinuing its use over time. What did they like most and least? What did they feel would improve it? Finally, we can measure how children's gains overall and on tests of specific skills varied with their usage of RoboTutor and the activities intended to teach those skills -- but how did efficacy vary across those activities, and why?

Analyze usage metrics from all five Finalists. Data logged differently by each Finalist may be problematic for anyone else to understand. However, XPRIZE would welcome this data and analyses of it to inform future development and deployment of intelligent tablet tutors.

Two decades ago, the U. S. President issued a challenge to “guarantee universal, excellent education for every child on our planet” [1]. We hope that the work reported here contributes to scaling up globally toward that eventual goal.

ACKNOWLEDGEMENTS

We thank the reviewers for their comments, and the many people who contributed to this project: A large, dedicated, almost entirely volunteer team developed RoboTutor. Mononito Goswami generated and refined the session and activity tables essential to data analysis. Liang Hao and Congkai Geng refined and applied our cognitive task analysis. The Simon Initiative at Carnegie Mellon University provided enthusiastic support. Colleagues provided generous statistical guidance. The Global Learning XPRIZE created the opportunity for this project and provided the test scores and social-emotional data analyzed in Study 1. Generous donors and the \$1M Finalist Award funded the development of RoboTutor. Last but certainly not least, the children who used RoboTutor generated the data analyzed in Study 2.

REFERENCES

- [1] Clinton, B. (1994). Public Papers of the Presidents of the United States, William J. Clinton. *United States: Office of the Federal Register, National Archives and Records Administration*.
- [2] Dangwal, R., & Gope, S.K. (2011). Indian adaptation of Motivated Strategy Learning Questionnaire, *International Journal of Education and Development using ICT* 7(3).
ijedict.dec.uwi.edu/viewarticle.php?id=1275.
- [3] King, S., Pressley, J., & Gove, A. (2019). Global Learning XPRIZE – Impact. Retrieved from data.xprize.org/global_learning_xprize/5cc0c16150ddf00be88dc6e1.
- [4] Kizilcec, R. F., & Cohen, G. L. (2017). Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *PNAS* 114(17), 4348-4353.
doi.org/10.1073/pnas.1611898114.
- [5] Korsah, G. A., Mostow, J., Dias, M. B., Sweet, T. M., Belousov, S. M., Dias, M. F., & Gong, H. (2010). Improving Child Literacy in Africa: Experiments with an Automated Reading Tutor. *Information Technologies and International Development* 6(2), 1-19. itidjournal.org/index.php/itid/article/view/517.
- [6] Ruipérez-Valiente, J. A., Halawa, S., Slama, R., & Reich, J. (2020). Using multi-platform learning analytics to compare regional and global MOOC learning in the Arab world. *Computers & Education* 146 (103776).
doi.org/10.1016/j.compedu.2019.103776.
- [7] Shukia, R., Katabaro, J., & Mgumia, J. (2019). Social Emotional Impact Assessment Draft Report. Report prepared for UNESCO, Tanzania.
- [8] Uchidiuno, J., Yarzebinski, E., Madaio, M., Maheshwari, N., Koedinger, K., & Ogan, A. (2018). Designing Appropriate Learning Technologies for School vs Home Settings in Tanzanian Rural Villages. In COMPASS '18: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, Article 9, pp. 1-11.
doi.org/10.1145/3209811.3209881.
- [9] Uchidiuno, J., Yarzebinski, E., Vargas-Vite, E., Koedinger, K., & Ogan, A. (2019). The Effectiveness of Publicly vs. Privately Assigned Group Leaders Among Learners in Rural Villages in Tanzania. In Lund, K., Niccolai, G. P., Lavoué, E., Gweon, C. H., & Baker, M. (Eds.), *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings*, 13th International Conference on Computer Supported Collaborative Learning (CSCL), Volume 1 (pp. 504-511). Lyon, France: International Society of the Learning Sciences.
repository.isls.org/handle/1/1610.
- [10] Vesselinov, R., & Grego, J. (2012). Duolingo Effectiveness Study Final Report.
static.duolingo.com/s3/DuolingoReport_Final.pdf.
- [11] Weber, F., & Bali, K. (2010). Enhancing ESL Education in India with a Reading Tutor that Listens, Proceedings of the First ACM Symposium on Computing for Development. Article 20, pp. 21-29. London, UK: ACM.
dl.acm.org/doi/10.1145/1926180.1926205.