A

Large-Scale Educational Campaigns

YUN-EN LIU, University of Washington
CHRISTY BALLWEBER, University of Washington
ELEANOR O'ROURKE, University of Washington
ERIC BUTLER, University of Washington
PHONRAPHEE THUMMAPHAN, University of Washington
ZORAN POPOVIĆ, University of Washington

Educational technology requires a delivery mechanism in order to scale. One method which has not yet seen widespread use is the educational campaign: large-scale, short-term events focused on a specific educational topic, such as the Hour of Code campaign. These are designed to generate media coverage and lend themselves nicely to collaborative or competitive goals, thus potentially leveraging social effects and community excitement to increase engagement and reach students who would otherwise not participate. In this paper, we present a case study of three such campaigns we ran to encourage students to play an algebra game, DragonBox Adaptive: the Washington, Norway, and Minnesota Algebra Challenges. We provide several design recommendations for future campaigns based on our experience, including the effects of different incentive schemes, the insertion of "tests" to fastforward students to levels of appropriate difficulty, and the strengths and weaknesses of campaigns as a method of collecting experimental data.

Categories and Subject Descriptors: K.3.1 [**Computer Uses in Education**]: Computer-assisted instruction

General Terms: Human Factors

Additional Key Words and Phrases: Learning at scale, educational games, large-scale educational campaigns

1. INTRODUCTION

Recent years have seen a surge of interest in educational technology as a means of delivering scalable, adaptive content to students, especially those who have not historically had access to high-quality resources [Guo and Reinecke 2014]. Massive open online courses (MOOCs) have been especially popular, but researchers have developed many other technologies that also qualify: intelligent tutoring systems [Anderson et al. 1995; Corbett et al. 1997; VanLehn 2006], adaptive curriculum content [Paramythis and Loidl-Reisinger 2004; Shute and Towle 2003; Wauters et al. 2010], and educational games [O'Neil et al. 2005; Mayo 2009] are all examples of educational technologies meant to both scale and improve student engagement and/or learning outcomes. The increasing availability of high-speed Internet has only accelerated this trend.

To encourage teachers and students to adopt and use educational systems, technologists must develop effective methods for packaging and delivering their content. In this paper, we study a new method of delivering educational content at scale: the state- or country-wide educational *campaign*. A campaign is a focused, widespread event where students from many schools engage in an educational activity over a short timespan. For example, in the Algebra Challenges we describe in this paper, students in grades K-12 were challenged to solve as many algebraic equations as possible in the educational game DragonBox Adaptive over a week-long period. And while we will focus primarily on educational technology, we believe that similar methods can be used by others to distribute other types of technology benefiting from concentrated collaborative or simultaneous use.

Campaigns are fundamentally different from typical methods used to deliver scalable educational technology. The simplest way to provide access to educational technology is by making it freely available online and allowing interested teachers to incorporate it into their classrooms, an approach taken by intelligent tutoring systems [Cor-

bett et al. 1997] or websites such as Khan Academy[1]. Another approach is to encourage participation by directly replacing the traditional classroom with an educational technology: MOOCs, for example, have attracted massive numbers of learners from diverse backgrounds with this approach [Breslow et al. 2013].

Unlike these other distribution mechanisms, campaigns explicitly use media and publicity to encourage the participation of schools and classrooms that might not otherwise seek out innovative educational technologies. They also typically involve substantial social interaction in the forms of both intra-classroom cooperation and inter-classroom competition. These factors make campaigns a unique method of distributing educational technologies, and have distinct advantages and disadvantages compared to more common distribution mechanisms.

There have been few large-scale campaigns involving educational technology; as a result, very little is known about their properties. In this paper, we explore data collected during three Algebra Challenges we conducted in Washington state, Norway, and Minnesota state. We describe how we iterated on the designs of our competitive incentives and our adaptation to player ability between each of the three Challenges, and analyze the resulting student outcomes. Based on analysis of the data from these Challenges, we provide a set of design recommendations for researchers and practitioners considering using similar campaigns as a method of delivering educational technology to many students. In particular, we focus on how student behavior in an educational game during our campaigns differs from behavior in games released to free online websites, how the incentive structures can improve student engagement but cause undesirable "gaming" behavior, the usefulness of inserting "pretest" levels to allow students to skip content they may already know, and advantages and disadvantages of using campaigns as an experimental platform to conduct educational research.

2. LARGE-SCALE CAMPAIGNS

In this work, we present a case study of three large-scale educational campaigns that we conducted on the topic of algebra. We define *campaigns* as events designed to encourage a large number of students to use an educational technology over a short period of time. Campaigns are designed to run at a state-wide or country-wide scale, attracting a large, diverse population of students. They are promoted through media and press releases, and possibly supported by visible figures such as politicians or industry leaders, as a way to recruit participants. Due to their limited timeframes, campaigns are designed to target a specific educational topic.

These types of educational campaigns have multiple goals. One goal is to generate awareness about both the educational technology being promoted and the organization that runs the event. Another goal is to motivate students and create excitement around the targeted educational topic. Educational campaigns are also designed to teach students during the event, and may have defined learning goals such as helping students achieve mastery on a certain concept or skill. Finally, large-scale campaigns can generate extensive data showing how a diverse population of students learns a topic or uses an educational technology. As a result, campaigns can be used as an opportunity to conduct controlled experiments on student learning.

The concept of using a short-term large-scale campaign for a targeted purpose is not new. A similar format has been used extensively to raise awareness about wide variety of topics, from eating disorders[2] to bullying[3]. Awareness campaigns typically

[1]http://www.khanacademy.org/
[2]http://www.http://nedawareness.org
[3]http://www.bullyingawarenessweek.org/

focus on information dissemination, with the goals of educating the public about a particular topic, encouraging behavior change, and raising money to support related research. One of the most well-known awareness campaigns is Breast Cancer Awareness Month[4], which was founded in 1985. Primarily focusing on awareness and fundraising, Breast Cancer Awareness Month includes walks for the cure, fundraising events, and extensive publicity and media events. As one example, the National Football League runs "A Crucial Catch Day" to promote breast cancer awareness by having players, coaches, and referees wear pink game apparel[5]. Studies show that Breast Cancer Awareness month helped increase the number of breast cancer diagnoses [Jacobsen and Jacobsen 2011], and a comparative study shows that the large scale and extensive publicity of Breast Cancer Awareness Month has made it more effective than the smaller-scale Bowel Cancer Awareness Month [Pullybank et al. 2002].

While health-related awareness campaigns are the most well-known campaigns, this model has been used to promote formal educational goals as well. Geography Awareness Week[6] uses the campaign model to encourage teachers to focus on geography-related lessons during one week in November. More recently, educational technology has begun to appear in campaigns. For example, approximately six months after our Washington Algebra Challenge, Code.org ran the *Hour of Code*[7] campaign and encouraged students across the United States to participate in an hour-long programming activity during Computer Science Education Week. Code.org partners also created 20 hours of interactive tutorial activities that teachers and students could choose from[8]. Unlike the Algebra Challenge campaigns we discuss in this work, the *Hour of Code* campaign focused on outreach instead of specific learning goals: on this measure they were highly successful, with 15 million students in 170 countries completing an hour of code during the one-week campaign[9].

Campaigns hold promise as a delivery vehicle for scalable educational content, but in this context are essentially unstudied. To the best of our knowledge, this work represents the first in-depth analysis of a large-scale campaign entered around educational technology. In this paper, we detail both the logistical and technical considerations of running this type of event, and discuss key findings from three campaigns of our own: the Washington, Norway, and Minnesota Algebra Challenges. We discuss lessons we learned through the process of organizing these campaigns, and present design recommendations for other researchers or organizations interested in conducting similar events.

3. EDUCATIONAL TECHNOLOGY DISTRIBUTION METHODS

Educational technology has been developed and studied for decades; however, the increasing availability of high-speed Internet has recently made it feasible for free educational systems to scale to massive numbers of students. There are a number of methods of providing access to educational technologies and encouraging students to use these types of resources. In this section, we describe existing methods of distributing content at scale, and discuss the advantages and disadvantages of each technique compared to the large-scale educational campaign method that we study in this work. It is important to note that we are more interested in the distribution mechanism than the content or technology delivered; the distinction has not always been made

[4]http://www.nationalbreastcancer.org/breast-cancer-awareness-month/
[5]http://www.nfl.com/pink
[6]http://www.geographyawarenessweek.org
[7]http://code.org/hourofcode
[8]http://codeorg.tumblr.com/post/73414288834/2013update
[9]http://codeorg.tumblr.com/post/70175643054/stats

clear in previous research, but in general most forms of educational technologies could be delivered through most of these distribution methods.

3.1. Purchased Software

The traditional method of distributing educational technology is as downloadable software that can be installed and run on computers in either a school or at home by purchase. This is the method that has historically been used to provide access to intelligent tutoring systems (ITSs). These advanced systems emulate one-on-one human tutoring using a cognitive model of the knowledge acquisition process [Anderson et al. 1995]. They are designed to help students as they work to master new problem-solving skills in a given domain. Most ITSs are intended to be used as a part of a course, and students are expected to have access to a textbook, an instructor, and possibly other resources [VanLehn 2006]. ITSs have been evaluated in the classroom with success; for example, students have shown significant learning gains when using ITSs as compared to paper-and-pencil activities in both algebra [Koedinger et al. 1997] and physics [Vanlehn et al. 2005]. These systems were originally developed before the advent of the Internet, and were designed to be run as desktop applications.

Schools and school districts also buy other types of educational software, such as reading programs, math games, or typing programs. Unfortunately, the market for such software and other educational tools can be hostile, especially to small groups and entrepreneurs [Foray 2011]. One of the primary reasons for this is that the effectiveness of popular software is often mixed, increasing suspicion among potential buyers: the popular algebra program *I Can Learn* was shown to raise test scores [Barrow et al. 2008], while the popular reading program *Fast ForWord* did not lead to increases in broader language acquisition or reading skills [Rouse and Krueger 2004]. Unfortunately, evaluating educational technology is expensive even for researchers, let alone educational companies, since they require developing relationships with decision makers at the school level, creating assessments, and organizing and analyzing the experiment [Chatterji and Jones 2012]. Similarly, selling and deploying such software can be challenging due to market fragmentation, differences in procurement methods between school districts, diverse curricula, and slow purchase cycles [Council of Economic Advisers 2011]. These and other factors, such as the propensity for running small resource-consuming pilots when start-up companies need larger deployments to remain viable, make it difficult for new entrants to create and market educational software [Berger and Stevenson 2007].

Campaigns serve a fundamentally different purpose from traditional methods of distributing software. They can avoid some of the problems plaguing the traditional educational software market, primarily due to their short duration and the fact that participation is free. This bypasses procurement problems that make selling software to idiosyncratic school districts difficult [Council of Economic Advisers 2011], and also involves a type of time-pressure and social incentive for schools to participate in an "exciting event" that is not usually present for purchased software. Campaigns should thus be considered a new tool for distributing educational technology, with different properties than existing methods. Intriguingly, their characteristics may make them useful for addressing some of the problems in current software markets: it may be possible to use them to run relatively inexpensive, large-scale studies or pilot tests to study their effectiveness or convince school districts to buy the underlying technology. We will touch on their viability as experimental vehicles later in this paper, but much research remains to be done.

3.2. Websites and Apps

Nowadays, one of the most common methods of releasing educational technology is through freely accessible webpages. Arguably the most popular of these websites is Khan Academy, which follows this model to deliver lectures and scaffolded exercises to students. We are not aware of any controlled studies of Khan Academy's educational effectiveness, though preliminary research by SRI International suggests that its use is associated with increased student test scores [Murphy et al. 2014]. The same study mentions that though Khan Academy is primarily used on an individual basis, its use as a supplemental resource in classrooms is increasing: the authors discovered that teachers use a variety of different methods to integrate Khan Academy materials within their classrooms.

Educational video games, which have experienced a surge of interest in recent years, are another type of technology that is typically delivered through free websites or app stores. Their popularity is in part due to their ability to engage players and motivate them to perform complex, time-consuming tasks [Gee 2008], and in part due to the observation that learning is an essential part of gameplay [Gee 2008; 2005]. In contrast, traditional education has been criticized for only successfully engaging a small portion of students [Skinner and Belmont 1993]; as a result, there is a growing interest in leveraging games to address the problem of student motivation in educational environments [Gee 2008; O'Neil et al. 2005; Mayo 2009].

Campaigns tend to be more top-down events in which schools and teachers encourage students to participate for a short period of time on a focused topic. Thus, like content available freely or cheaply online, they are likely to be used in a supplemental fashion. Unlike online content, however, campaigns are also designed to increase student motivation and excitement through social pressure and collaborative or competitive rewards for participation. That is, they "push" content to students more, compared to the usual "pull" model of a student choosing to visit a website or download an app to view content.

3.3. Structured Online Courses

Over the past few years, Massively Open Online Courses (MOOCs) have become the central focus of discussions about learning at scale. These courses are designed to replace traditional classrooms with massive online classrooms, hopefully enabling the participation of diverse groups of students who have not previously had access to high-quality post-secondary education [Breslow et al. 2013]. Hundreds of thousands of students have earned certificates through MOOCs in diverse topics such as music, systems biology, and computer science [Lewin 2012]. In contrast, campaigns are much more limited in scope, and are not designed to replace traditional classes in any sense. Thus they do not have lecturers or staff, as MOOCs do, and do not necessarily need to assign grades or credit for work completed (though they can do so if the campaign is run as a competition). More fundamentally, campaigns are by their nature exciting community-wide events, and one of their primary goals is to generate student interest in a particular topic. MOOCs are less concerned with this sort of outreach, and more about teaching already-interested students deep knowledge about a topic.

4. THE ALGEBRA CHALLENGES

This paper is primarily a case study about three campaigns we ran: the Washington, Norway, and Minnesota Algebra Challenges. These campaigns were run sequentially, in order to give us a chance to make changes and improve outcomes from one to the next. We will first describe the educational game delivered to students through the campaign, DragonBox Adaptive, then give some basic information about the cam-

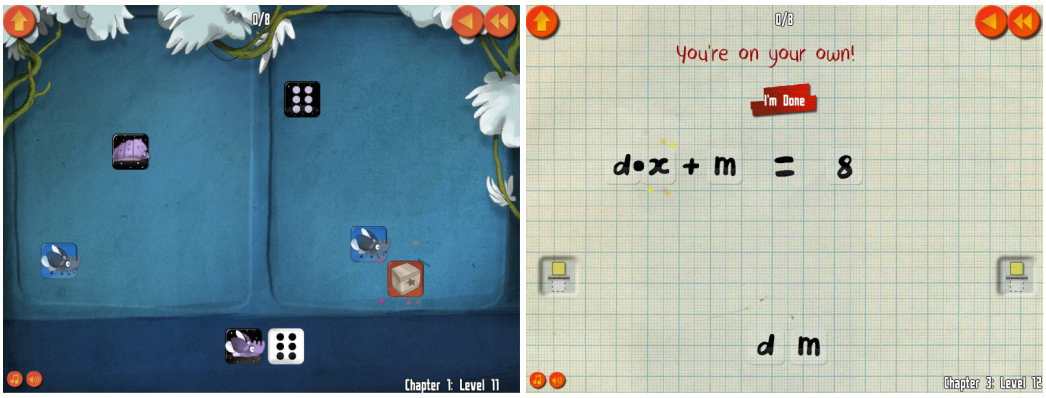
paigns themselves and how their structures differed. Later, we will take a closer look at the data we gathered during the Algebra Challenges and offer design advice to other researchers considering such campaigns, either as ways to deliver educational content or as experimental platforms for studying educational interventions.

4.1. DragonBox Adaptive

For the educational content delivered through our Challenges, we used a game called DragonBox Adaptive [Center For Game Science 2014], evolved from the original game DragonBox [We Want To Know 2014]. The game is designed to teach algebra for all students in the K-12 spectrum, and can be seen in Figure 1. Each level represents an algebra equation to solve, with left and right sides filled with cards that represent numbers and variables. The key card is the Dragon Box, which must be isolated on one side for the player to win (much like X in an algebraic equation). Players can perform algebraic operations by simplifying cards in the equation area (e.g., '−a+a' → '0'), eliminating identities (e.g. '0' → ' '), or using cards in the deck at the bottom of the screen to add, divide, or multiply both sides.

An important concept in the version of DragonBox Adaptive we used in the Algebra Challenges is *mastery*, one of our outcome measures. Students demonstrated mastery by successfully answering a series of three test levels designed to look very similar to standard paper-and-pencil algebra problems. An example of one of these tests can be seen in Figure 1(b). Test levels offered no hints, did not force students to keep both sides balanced, and allowed students to submit answers at any time. Furthermore, students had to submit correct answers to three tests in a row under a certain time limit in order to achieve mastery. We will see in Section 5.4 that students were largely unable to pass these tests when randomly given out early in the game, but had much better success rates when given sufficient practice; thus, at the very least, students in the Challenges improved at solving these mastery test levels.

As the name implies, DragonBox Adaptive differs from its parent game, DragonBox. The primary difference is adaptivity: depending on how they performed on embedded



(a) Simple (b) Mastery

Fig. 1: Screenshots of DragonBox Adaptive, ©(2013) UW Center for Game Science / WeWantToKnow AS [We Want To Know 2014]. Figure 1(a) shows an early level of the game with the equation a−b=−6+a+x. The DragonBox, on the bottom right, must be isolated on one side to pass the level. The game increases in complexity and gradually begins to look more like standard algebra, as can be seen in one of the mastery test levels in Figure 1(b).

Location	# Users	Time (sec.)	Mastery (all/1.5h)	Partner	Start	End
Washington	4200	4199	52%/96%	Technology Alliance	2013-06-03	2013-06-08
Norway	36100	7375	65%/96%	We Want to Know	2014-01-12	2014-01-17
Minnesota	6900	3698	48%/96%	Technology and Information Education Services	2014-02-03	2014-07-01

Table I: Some basic statistics about the three campaigns we ran. Note that the bulk of Minnesota players played in the first week. Time is the average length of time players were actively playing. The two mastery rates are the overall rate, and rate among students who actively played at least 1.5 hours (the length of time we asked teachers to devote). More in-depth discussion and statistical analysis will be given in Section 5.

assessments, students were given different sets of dynamically-generated problems for additional practice. These assessments measured overall success, time to completion, and number of required actions to make these decisions. Furthermore, there were two types of assessment levels: those which looked and functioned like normal levels with in-game scaffolding, and the test levels mentioned above that looked like algebra problems and had no scaffolding.

All three Challenges gave mastery test levels at a particular point in the progression (about 50 levels in), and if the player failed the test they were sent backwards in the progression to play a few more levels before trying the test again. This additional practice usually amounted to about three levels, though occasionally could be more if a student performed very poorly. Norway and Minnesota were enhanced with a second set of even more difficult mastery tests; again, if students failed the tests we generated new practice problems for them.

All three Challenges also made use of the assessments that looked like regular game levels. In Washington, these levels were restricted to later on in the progression just before the mastery tests, so that all their extra practice tended to be extra levels involving relatively difficult concepts. In Norway and Minnesota we added more of these assessments for even basic concepts, so that students who needed additional practice early had the opportunity to play extra levels. In all cases, we used a "generative" adaptivity where new problems were generated for each student depending on which assessment they failed; this is in contrast with the common tutoring system strategy of drawing from an existing and finite pool of problems.

4.2. Algebra Challenge overview

We ran three Challenges in sequence: Washington, Norway, and Minnesota. Washington and Norway were designed to run for one week, and while we set a long cutoff date for Minnesota, the bulk of the activity occurred during the first week there, as well. To recruit teacher and student participants, we partnered with different organizations with existing connections to schools: we did not pay them, and indeed in some cases they paid us. In general, we found that it was quite easy to find teacher organizations and educational technology promotion groups eager to run these types of campaigns. These groups contacted all interested teachers through mailing lists, bulletins, social media, or through word-of-mouth. We did not place any grade restrictions on signups,

Fig. 2: The Algebra Challenges included a great deal of supporting infrastructure, including the "Challenge Dashboard." This website functioned as a leaderboard that teachers and administrators could use to track mastery, number of equations solved, and play time at the class, school, or district level. Furthermore, teachers were able to track student progress within their own classroom.

so we received a wide spread of students. Future campaigns with specific pedagogical goals may wish to target content towards a particular grade; we do not know how easy or difficult this would be, but imagine that schools would be less eager to participate if only a fraction of students are allowed to participate. We also had some media coverage and promotion by political figures: for example, we received a letter of support from Washington Governor Jay Inslee, and the Norwegian Digital Learning Arena tweeted a video of the Norwegian Prime Minister, Erna Solberg, playing DragonBox Adaptive as part of the Challenge. This support may have lent legitimacy to our efforts and increased participation, though the strength of this effect is difficult to measure.

Before examining more specific information about the Challenges, what makes campaigns successful overall? We are not equipped to answer this question with certainty, but can speculate. Certainly the Algebra Challenges had a great deal of supporting infrastructure, such as collaborative or competitive goals and websites to track student or class progress, registration and organization pages, and server architecture. The Challenge Dashboard seen in Figure 2 seemed quite popular among teachers, for instance. Yet at the end of the day the primary consideration for any campaign is how many schools, teachers, and ultimately students choose to participate.

Our sense from communications with politicians, technology advocacy groups, administrators, and teachers, is that everyone benefited in different ways. Politicians could generate positive press with their association with an innovative, large-scale educational effort. Advocacy groups could both tout the benefits of educational technology and increase their own influence by adding a major campaign to their list of accomplishments. Teachers and schools seemed excited to generate enthusiasm among students by being part of a "big" event involving an educational game. For everyone, the Challenges inspired a sense of being a part of a cutting-edge, important project run by a prestigious research institution. In addition, even though the content was experimental and not guaranteed to deliver learning gains, the limited time commitment and use of a modified popular algebra game, which students would enjoy regardless, meant there were few potential downsides. If true, this argues in favor of both short-duration and game-based campaigns when content is relatively untested.

Another important question to ask is the following: how easy would it be for other groups to run such campaigns? This is very difficult for us to answer. From our perspective, running the Challenges required some amount of effort and coordination, but was not tremendously difficult. The most important part, teacher recruitment, was mostly handled by other parties, such as the press generated by politicians or mailing lists of the technology groups we partnered with. At the same time, our group is relatively well-known for other projects, has well-connected funders, and was using an advanced version of a popular existing math game. If none of these factors had been in play, we may not have been able to convince these groups that the effort of running such events would have been worthwhile. We doubt, for example, that a research team consisting of five people with no existing infrastructure, connections, or credentials would be able to run a successful campaign. Speculation aside, the best way to answer this question is for other research groups to attempt to run their own campaigns and report their experiences.

Now that we have covered our general observations about running campaigns, some basic statistics about the Challenges themselves can be seen in Table I. As a reminder, "Mastery" refers to the ability of players to pass three difficult test levels in a row, under a certain time limit. Since many of these players failed to reach mastery because they did not play long enough, we also include mastery rates for students who played at least the 1.5 hours we requested of students. These are quite high, but we note that this is no guarantee that if students who stopped were compelled to play longer they would reach similar rates of mastery. Another important point is that higher play times are not necessarily better: they could be indicative of more engagement, but could also reflect an excess of time spent on easy problems. For example, the average time to mastery in the Washington, Norway, and Minnesota Challenges was 2626, 2820, and 2515 seconds, respectively: thus, the Minnesota Challenge may have been more efficient in some sense. That being said, the data in the table suggest that Norway was a clear outlier both in terms of scope and player behavior. There are several possible explanations for the difference in time played and mastery rate in Norway as compared to Washington and Minnesota. Since this is a case study and many variables changed each campaign, we cannot definitively answer this question, though we can still examine individual hypotheses.

One possible explanation is that Norway students tended to be older, as can be seen in Figure 3. We can test this hypothesis by running an ANCOVA on Time Played (in seconds) with independent variable Location and covariate Grade. There is a significant difference between the three Algebra Challenges ($F(2, 47197) = 250.154, p < .000$), while the association between Time Played and Grade is not significant ($F(1, 47197) = 1.361, p = .243$), suggesting that the population age is not the sole driver of the increase. Only 2.8% of the variance is accounted for by Location, controlling for Grade

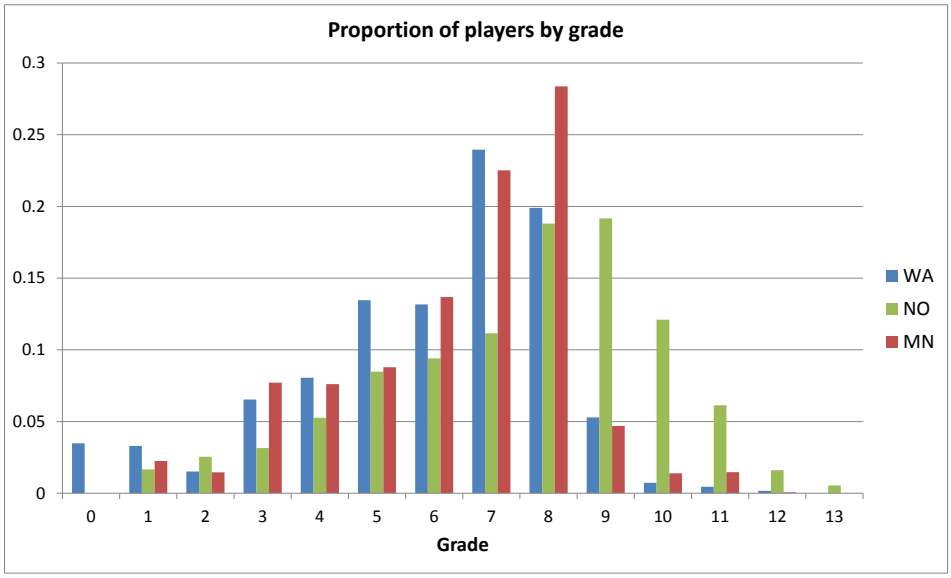


Fig. 3: The average grade levels for Washington, Norway, and Minnesota are 5.9, 7.2, and 6.4, respectively. An ANOVA shows these differences are significant ($F(2, 47198) = 1335.626, p < .000$). The mean Norway grade is higher than Minnesota's ($p < .000$) and Washington's ($p < .000$), and the mean Minnesota grade is higher than Washington's ($p < .000$). This is a potential confound to keep in mind.

($R^2 = .028$), so other unmeasured factors may be responsible. To complete the analysis, we also find that Norway students (Mean=7375 seconds) play statistically significantly longer than Washington (Mean=4199 seconds) and Minnesota students (M=3698 seconds) at the .01 level ($p < .000$), while the difference in Time Played between Washington and Minnesota is not significant at the .05 level ($p = .058$).

Another intriguing possibility has to do with how we encouraged participation. The incentive scheme in Norway was different than the one used in Washington, and potentially more motivating; Minnesota had no incentive scheme at all. Could the change in incentive structure explain why students in Norway played longer than those in Washington and Minnesota? We will analyze the data in light of this hypothesis in Section 5.2.

5. LESSONS

Campaigns seem like a promising way to deliver certain types of educational technology to many students. To help other researchers or organizations who wish to run similar events, we will share some of our observations from the Algebra Challenges. We will discuss potential differences between student data generated by campaigns compared to free game websites, the effects of different incentives on student behavior, what factors we observed were important to account for if attempting to run a randomized experiment in a campaign, and the results of one randomized study we ran in the Minnesota Algebra Challenge when attempting to fastforward students through content using early mastery tests.

5.1. Comparison to Online Educational Game Releases

To better understand how campaigns function as a mechanism for distributing educational technology and encouraging participation from a broad population of students, we were interested in comparing the DragonBox data collected during our campaigns to data collected from educational games released through online game portals. This type of analysis allows us to highlight features of campaigns that make them different from standard distribution methods, and define the properties that educational systems must have to be appropriate for distribution through a large-scale campaign.

DragonBox Adaptive has not been released on any free online game portals and the maker of the original game DragonBox does not track engagement statistics, so we are unable to directly compare student behavior in this game across distribution methods. However, many other educational games have been released online, so we can compare high-level features of the data collected with these games to data collected with DragonBox Adaptive through our campaigns. This will give us an initial understanding of the differences between free game portals and campaigns as a distribution method for educational games.

In this analysis, we compare the Algebra Challenge data to data collected with games released to the popular educational website BrainPOP[10]. BrainPOP provides an educational games portal designed for use in the classroom, where games can be freely accessed at any time. We look at three popular BrainPOP games: Refraction[11], which teaches fraction concepts through splitting lasers, and Treefrog Treasure[12] and Battleship Numberline[13], which both teach number line concepts. All three games were developed for research purposes, and student behavior in each game has been studied extensively [Lomas et al. 2013; O'Rourke et al. 2013].

Students who participated in the Algebra Challenge played DragonBox for much longer periods of time on average than students played the other three educational games on BrainPOP. In the Washington, Norway, and Minnesota Challenges, students played for a median time of 54 minutes, 82 minutes, and 47 minutes respectively. In comparison, published literature suggests students on BrainPOP play Refraction for a median of 3 minutes [O'Rourke et al. 2013], Treefrog Treasure for a median of 8 minutes [O'Rourke et al. 2013], and Battleship Numberline for a median of 2-3 minutes [Lomas et al. 2013] (reported numbers in this paper are reaction times, so true time played is higher).

We caution that these numbers cannot be directly compared, given that the underlying games are different. Furthermore, it is not easy to track students on free game websites across different sessions without login information, so most research on free online games measures time played during the first session only. We have no such problem tracking across sessions in our Challenges and so can measure complete time played. With these caveats in mind, there are good reasons to believe that students may play more seriously in campaigns compared to online games. Teachers may plan for students to play the educational game for a set period of time during class when participating in a campaign, for example. There may also be effects from collaborative goals or rewards for campaign winners, though we note that students still played quite long in the Minnesota Challenge, in which there were no explicit incentives. Whether or not campaigns truly promote greater engagement than online distribution methods, and exactly how students use the technology differently depending on the distribution method, are questions we leave for future research.

[10]http://www.brainpop.com/
[11]http://www.brainpop.com/games/refraction/
[12]http://www.brainpop.com/games/treefrogtreasure/
[13]http://www.brainpop.com/games/battleshipnumberline/

Location	Collaborative	Competitive	Prizes
Washington	250,000 equations, Class mastery	Class mastery	One tablet per grade per class size
Norway	400,000 equations, Class mastery	Equations solved	One tablet per student of the winning class
Minnesota	250,000 equations, Class mastery	None	None

Table II: Incentive structures in the Algebra Challenges. The collaborative goal was a target number of equations to be solved across all students participating in the campaign, and progress towards the goal was listed on the campaign website. The competitive goal was a reward given to classes that had the highest mastery rates (in Washington) or the most equations solved (in Norway).

Another difference between the data collected during the Algebra Challenges and through online game portals is the type of demographic data collected. During the Challenges, we were able to collect gender, grade, and school information about students, given by teachers when they registered their classes. This data is likely to be reliable because it is entered by teachers rather than students, in stark contrast to the data collected through online educational game portals. O'Rourke et al. note that it is challenging to collect demographic data from students who play games on BrainPOP [O'Rourke et al. 2014], and demographic information is not presented in any of the work studying Refraction, Treefrog Treasure, or Battleship Numberline [Lomas et al. 2013; O'Rourke et al. 2013; Andersen et al. 2011]. We have tried collecting demographic information in educational games released online through embedded, skippable questionnaires, but have found that these surveys cause the majority of players to quit; more complex methods, such as giving players the opportunity to enter demographic information to compare their responses to similar players' responses [Reinecke and Gajos 2014], may be required.

These comparisons show that it is easier to collect "complete" data through campaigns than free online games. That being said, campaigns are also more expensive to conduct. By our estimates, each campaign cost around $45,000 to conduct, while releasing a game to an online portal is free. However beyond just the cost, there are other features of campaigns that educational technologists should consider. Since participants play for a very long time on average, the content of the educational technology being distributed must be extensive enough to support multiple hours of play. Depending on the stated goals of the campaign, near-infinite content may be desirable. In the Norway Challenge, for example, classes were rewarded for completing the largest number of equations: this meant that DragonBox Adaptive needed to support an infinite amount of play time to fairly support this competition. We accomplished this by procedurally generating levels from templates, to ensure that students could continue playing for as long as they wanted; indeed, one student in Norway played almost 23,000 levels.

Even without such outliers, the fastest 5% of players in the Norway Challenge achieved mastery in 64 levels, while the slowest 5% of players achieved mastery in 224 levels. If such content generation or adaptivity is not possible, organizers of campaigns may wish to promote completion of the content instead of duration or volume of interaction; however, our experience suggests that having generative adaptivity may be useful to accommodate the wide spread of student abilities.

5.2. Incentives

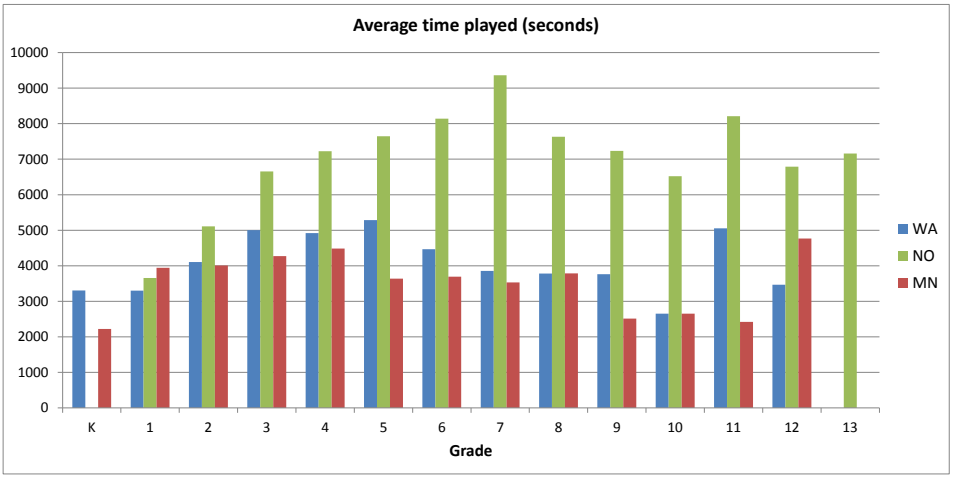
Educational technology is only valuable if students feel motivated to use it. Incentives for using a given system can vary widely. Students may have a desire to learn, may be seeking accreditation in the form of a badge or certificate, may enjoy the activity, or may be encouraged or pressured to participate by teachers, parents, or peers. Luckily, campaigns are uniquely suited to take advantage of many of these incentives. In our three Algebra Challenges, we used a mix of incentives not frequently found in other methods of delivering educational content at scale. In each Challenge, we set a collaborative goal of solving a certain number of equations across all participants. We also experimented with different types of competitive goals in the Washington and Norway Challenges as a way to leverage peer excitement and peer pressure at the classroom level to increase student engagement.

The incentives we used for each campaign can be seen in Table II. There are a number of differences between the competitive incentives used in Washington and Norway. In Washington, we incentivized mastery rate, or students' ability to reach a certain point in the game. We chose to reward mastery because the educational goal of the Challenge was to help students master algebra concepts. We awarded one tablet to the classes that had the highest overall mastery rate at each grade level. In addition, some classes had many more students participate than others, and it seemed unfair to compare large classes to small ones. We therefore awarded prizes to one class of each size (extra small, small, medium, or large) at each grade level, for a total of four winning classes in each of the grades K through 12.

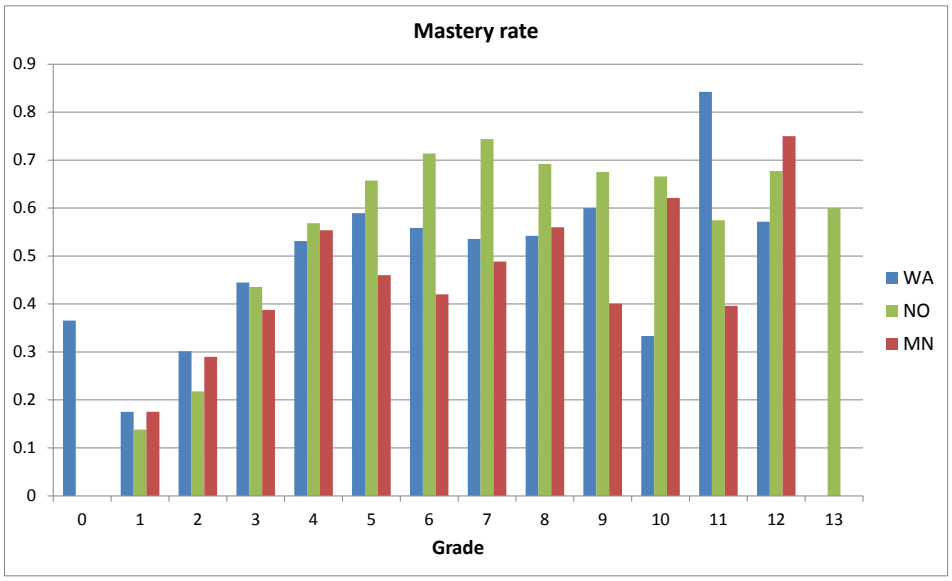
In Norway, the sponsors requested we change the incentive structure to instead reward the total number of equations solved by each class. The goal of this modification was to reward behaviors that students could see and directly control. It is very clear when a student solves an equation, because this is equivalent to completing a level in the game; mastery, on the other hand, is a much more nebulous concept that was more difficult for students to grasp and work towards. This idea is similar to one proposed by Fryer, who found that giving students financial incentives for reading books significantly improved outcomes, while providing the same incentives for either increasing classroom grades of state test scores produced no significant improvement [Fryer 2011]. We therefore hypothesized that a similarly concrete incentive structure might be more effective in our campaigns. We further decided to reward only the single class that solved the most equations, and we gave each student in the winning class a tablet; this was designed to increase the personal incentive for students to play.

As seen previously, the descriptive statistics in Table I and the ANCOVA results from Section 4.2 immediately suggest two possible conclusions. First, we found no significant change in how long students played between the Washington and Minnesota Algebra Challenges even after controlling for grade. Given that we did not advertise any incentives in Minnesota at all, this calls into question the usefulness of the Washington incentive model. Second, students in Norway put much more time into the Challenge: they played for a much longer period of time on average, even when controlling for grade. This suggests that the more personalized and direct incentive structure may have been effective at influencing student behavior.

We caution that the higher time played and mastery rate during the Norway Challenge could be the result of any number of factors. Perhaps the schools involved in the campaign were richer, or maybe DragonBox is better-known in Norway. Unfortunately, it is difficult to study the impact of different incentive schemes, because it is challenging to conduct randomized experiments in campaigns. We cannot run two simultaneous campaigns in the same schools and same location with different incen-



(a) Time



(b) Mastery

Fig. 4: Differences in player outcomes persist even when accounting for the fact that Norway students were older.

tive schemes, for example. Thus all analysis that follows should be taken primarily as suggestions for future research directions, possibly in a more controlled environment.

Caveats aside, the time played and mastery rate in the Norway Challenge are encouraging, and suggest that our equation-based competitive incentive structure may have been effective. For example, in Norway approximately 44% of levels played were played outside of regular school hours; in Washington, this figure was a still-respectable but lower 21%. However, the effects were not all positive. We also uncovered evidence of undesirable behavior while performing statistics to find the winning classroom. In the Washington and Minnesota Challenges, the maximum number of

Location	Egregious skipping	Any skipping
Washington	0 students (0%)	7 students (0.2%)
Norway	2111 students (5.8%)	10133 students (28%)
Minnesota	0 students (0%)	0 students (0%)

Table III: Occurrence of intentional failure by "skipping" a problem (submitting an answer without making any moves) This is an example of undesirable "gaming the system" behavior, potentially a side effect of the Norway incentive scheme rewarding total levels completed and giving tablets to each student of the winning class.

levels played by any single student were 1,634 and 3,000, respectively. In Norway, however, 90 players played more than 3,000 levels, and the maximum number of levels played was 22,960. This far exceeded the amount of practice we desired from students, and strongly affected the total class equation counts used to determine the winner.

When we examined the data more closely, we discovered that some players were "gaming the system." There were two different types of gaming behavior that we observed. The more mild form involved players submitting incorrect answers to game levels. This behavior would cause the game to restart the level with a potentially easier format, allowing students to boost their numbers of equations solved by completing many easy levels. The more egregious form of gaming behavior involved players deliberately failing a mastery test, causing them to be sent back to easier levels that could be solved very quickly before reaching (and failing) another mastery test. In other words, these players used the ability to replay easier levels to artificially inflate the number of equations solved.

A natural question is whether this type of gaming behavior occurred at the same rate in all three Challenges, or whether the rate was elevated during the Norway Challenge. A reasonable proxy for intentionally giving up on a regular game level (mild gaming) or a mastery test (egregious gaming), is to measure how often participants submit an answer without having made a single move. Table III shows the number and percentage of players who engaged in this behavior at least once for arbitrary levels or mastery levels. Given that almost no intentional failure occurred in either Washington or Minnesota, it seems highly likely that this undesirable gaming behavior occurred as a direct result of the different incentive structure used in the Norway Challenge.

While these results are suggestive, more research is needed to understand how best to incentivize participation in these types of campaigns. We were not able to determine from this analysis whether rewarding individual over joint achievement was more important, or rewarding total levels completed rather than mastery rate was more important. Furthermore, while a sizable portion of students in the Norway Challenge engaged in gaming behavior, the majority did not. It is possible that there is a way to retain the gains in time played and mastery for the majority of the population, while reducing opportunities to game the system. Can the game rules be structured to continue to offer additional support for struggling players while punishing intentional player failure? Or can we detect gaming behavior when it happens, as has been tried in Intelligent Tutoring System [Baker et al. 2004]? Regardless of the answers to these questions, the results from our analysis suggest that some types of incentives can affect student engagement and participation in campaigns; thus, these incentives must be designed carefully to encourage desirable behaviors.

5.3. Campaigns as an experimental platform

Educational software is increasingly being used to run experiments, as has been the case in the e-commerce and game industries for a long time [Kohavi et al. 2007]. No-

tably, several studies using data drawn from educational games have been published in the past few years (ex. [Lomas et al. 2013; Liu et al. 2014]). Previous researchers have considered using online experiments, combined with more traditional sources, to construct new types of experiment frameworks [Stamper et al. 2012]. Challenges could conceivably be used for this purpose as well, but have their own unique set of strengths and weaknesses relative to other content delivery mechanisms.

Clear strengths of campaigns as experimental platforms are as follows:

— Relatively high rate of data (36,000 players in Norway over 5 days).
— Enhanced persistence of students compared to games released on free educational websites.
— Access to some demographic information about students, entered by teachers (note that this mitigates a key drawback of internet data, which generally does not contain rich data [Stamper et al. 2012]).
— Relative ease of some randomization and data collection (updating a webpage and consulting a database, respectively).
— Potential ability to run social or multiplayer interventions, since it is very likely that students will be participating as part of a group.

Clear weaknesses of campaigns as experimental platforms are as follows:

— When cast as competitions, unfairness and negative press can become problematic. It would have been nearly impossible to run any kind of substantial randomization in Norway, for example, as running a competition requires that every class be using the same game and level progression to fairly compare classes against each other. This is much less of an issue if the only stated goals are collaborative: this was the case in Minnesota, and as we will discuss in Section 5.4 we were able to run a randomized experiment with fastforward mastery test levels.
— Violation of independence of samples, since students may play together. Precisely measuring how often this occurs is difficult, but two statistics are indicative: in the Washington Challenge, 80% of levels were played during standard school hours (8:00-15:00), and 98.5% of levels were played within one minute of another student from the same class playing a level. This problem is particularly troublesome if different players in the same class are put into different experimental conditions, since students may notice that they are not performing the same task.

Other characteristics of campaigns are as follows:

— Each Challenge cost approximately $40,000 in salary and $4,000 in server costs, not including the development of the logging infrastructure and Dragonbox Adaptive.
— Our campaigns only run a few days: the upside is that there is less time for other factors to interfere (such as learning outside of school), but the downside is that it is difficult to measure retention over long timescales or adjust for any problems that occur in the middle of the campaign.
— It may be possible to interview individual students, since the participating schools and classes are known. We did not do so during our campaigns, as our primary focus was on running them smoothly and with maximum participation; we suspect interviews would have been possible, but coordinating permissions and timing with instructors might have been difficult. It will be important to investigate the ease of collection and usefulness of qualitative data in future campaigns.

To mitigate the effect of unfairness while retaining the highly motivating competitive incentives, one possible solution might be to run competitions at the school level instead of globally. Thus school A and school B might have their own distinct prizes to award, and could be safely assigned to different experimental conditions. For that

matter, the same could be done at the teacher level, allowing randomization between teachers within the same school. This would be ideal, but only if the school or teacher have little effect on student behavior so that students in different experimental conditions would be a priori similar. Likewise, we might also wonder whether other student characteristics, such as gender and socioeconomic status, seem to affect their behavior. To answer these questions, we will attempt to model the effects of school, teacher, grade, gender, and socioeconomic status in the Algebra Challenges. The results are useful in understanding how researchers should try to randomize students into conditions, and which covariates may be important to pay attention to in the process.

5.3.1. Hierarchical Linear Model. How much of an effect on total time played do school and teacher have in our dataset? To answer this question, we used Hierarchical Linear Modeling (HLM), a complex form of ordinary least squares (OLS) regression, which analyzes the variance in the dependent variable when predictors are at varying levels: e.g., students in a classroom share variance as a function of their common teacher and common classroom. While this data could be analyzed ignoring the nested structure of the data, for example using fixed parameter simple linear regression, this approach would be insufficient due to its disregard for the shared variance [Woltman et al. 2012], aggregation bias and misestimated precision [Raudenbush 1988].

More precisely, a 3-level Hierarchical Linear Model was employed to analyze how much of an effect schools and teachers had on the amount of time played, in seconds, during each of the three Algebra Challenges. Student was Level-1 of the model, Teacher Level-2, and School Level-3. Our dependent variable for these models is Total Time Played, measured in seconds. Our independent variables are Grade, Free and Reduced Lunch (Washington and Minnesota only), and Gender (Norway only). Free and reduced Lunch is the percentage of students in a school that participate in the Free and Reduced Lunch program, a common socioeconomic status indicator in education research, and was collected from school reports posted online.

Due to the fact that the data available for each Campaign varies, our models for Washington and Minnesota are slightly different from the Norway model. For Washington and Minnesota we added Grade as a predictor in the Student Level and Free and Reduced Lunch as a predictor of the School Level. Norway had Gender data for a portion of their students, so we added Gender as a predictor in the Student Level along with Grade; however, Free and Reduced Lunch data was not readily available for Norway, and so was not included as a predictor in the School level.

Results of the HLM analyses are provided in Table IV; the findings vary by location. In Washington, the effect of Grade on Time Played was not significant, while the effect of Free and Reduced Lunch was statistically significant at the .01 level ($t = -4.125, p < .000$). This means that for Washington, the Grade level does not impact the length of time a student plays the game, and the higher a school's free and reduced lunch participation rate, the less length of time students from that school play. In Minnesota, the effect of Grade does have a statistically significant impact at the .05 level ($t = -2.265, p < .024$), where higher grade levels play less time than lower ones, while the impact of Free and Reduced Lunch does not have a statistically significant impact. In Norway, while Grade is not significant, Gender is at .05 level ($t = -0.571, p = .568, t = 2.270, p = .023$, respectively), with female students playing longer than male students.

With regards to the influences of teachers and schools, the analyses showed that the proportion of variance within teacher is more than 33% for all locations (Washington 45.60%, Minnesota 42.33%, Norway 33.82%). As compared to variance within teacher, the proportion of variance among teacher within schools is lower for Minnesota (33.24%), but marginally higher in Norway (38.51%) and Washington (48.03%). In contrast, the proportion of variance among schools in Washington (6.38%) is much

Fixed Effect	Coefficient	Standard Error	DF	t-ratio
WA				
Intercept	5402.42	539.13	61	10.021**
Grade	-134.57	112.68	2281	-1.194
Free/reduced lunch	-3629.95	880.06	61	-4.125**
MN				
Intercept	4006.83	791.89	53	5.060**
Grade	-70.40	31.08	3801	-2.265*
Free/reduced lunch	-1395.19	1742.95	53	-0.800
Norway				
Intercept	9163.89	541.60	287	16.920**
Grade	-105.84	185.50	13464	-0.571
Gender	444.11	195.63	13464	2.270*

Table IV: *$p < 0.05$, **$p < 0.01$. 3-level HLM results, with Student nested in Teacher nested in School, intercepts and standard errors given in seconds. Highly significant intercepts suggest that the school and teacher impact how long students participated, which may be useful for future experimental designs using campaign data.

lower, while the proportion of variance among schools in Minnesota (24.44%) and Norway (27.68%) are still high. Thus teacher and school account for a great deal of variance and appear to impact students' Time Played, and potentially the school matters less for explaining student engagement in Washington.

5.3.2. Discussion. Regardless of location, we found that teacher and school were highly significant factors on how long students play. This means, for researchers who wish to run randomized experiments using campaigns, randomizing within schools or even within teachers is ideal. If the researcher wishes to use competitive incentives and the conditions are very different, then this may not be possible. Further research is required to understand how characteristics of schools and teachers predict student behavior, so that conditions can be balanced as evenly as possible. Furthermore, the characteristics that are important may vary by location: Free and reduced lunch rates, for example, are significant predictors in Washington but not Minnesota.

At the student level, we discovered that gender was a significant predictor in Norway, with females playing longer than males. In Washington and Minnesota, grade was either not significant or had a small effect; this is not to say that grade level is not important, but rather that once the school and teacher are known the grade level has little effect. Grade was both significant and had a larger impact in Norway, with higher grades playing less time once the school and teacher were known. Why exactly this is and why the results are different between the three locations deserves further research; we note in passing that primary school in Norway typically spans 7 years, while primary school in the U.S. typically spans 5 years, so that there may be more opportunity for grade to make a difference within Norwegian schools.

Besides providing guidance to future researchers wanting to use campaigns as a source of experimental data, our results also suggest potential improvements to the design of campaigns themselves. We motivated this analysis by asking whether or not researchers could avoid the need to randomize within schools or teachers, in order to take advantage of competitive incentives without causing unfairness. Ideally, though, we could design other non-competitive incentives with equal or greater effects than

prizes for top classes, in which case this may no longer be a concern. Furthermore, the significant result of gender in Norway and free and reduced lunch percentages in Washington schools suggest that campaigns may not be "fair." In particular, one of the essential motivations for scalable learning technologies is the ability to reach otherwise disadvantaged students: unfortunately, our results suggest that students at Washington schools of low socioeconomic status participate for less time. Future campaigns may thus wish to provide additional outreach or support to these disadvantaged students to ensure they have the chance of equal participation.

5.4. Fastforwarding

The population of students who participated in our Algebra Challenges were very diverse in both their ages and their incoming understanding of algebra: looking at the wide grade spread in Figure 3 should be evidence of this. Another piece of evidence is that we saw a wide range of times to achieve mastery in the Washington and Norway Challenges: for example, the 10th, 50th and 90th percentiles of time to achieve mastery in Norway were 1322 seconds, 2545 seconds, and 5070 seconds respectively.

Given that educational campaigns draw many types of students, and that students engage the material for a fair amount of time, some form of adaptivity is likely to be important to prevent large groups of students from being bored or frustrated. One possible method is to give a test early on, and fastforward the student to a more difficult part of the progression if the test is passed. We were able to study the effects of such a "fastforward" strategy of adaptivity in Minnesota by giving random mastery tests, and will analyze the results in this section.

We provided two types of adaptivity in DragonBox Adaptive, as was described in Section 4.1. These two types of adaptivity were both designed to provide *additional* practice for players who were struggling. However, players who were already adept at solving algebraic equations could become bored while working their way through the 52 minimum levels required to reach the mastery tests: that is, the progression might have been too slow for players who already understood algebra. An important component of games and other optional educational technologies is that they must be engaging, or else students may simply stop and do something else. This led us to study a third type of adaptivity in the Minnesota Challenge: with 0.65% probability, the mastery tests would be given after any level before the normal mastery test location. If passed, the player would be fastforwarded to the progression after the test; if failed, the player would be returned to the original point and continued to play as normal (possibly receiving more fastforward tests later). This had the potential to greatly reduce the number of levels required to achieve mastery.

Unlike most of the other data analysis in this paper, which is confounded by simultaneous changes in incentives, population, and level structure, these fastforward tests can be considered a randomized experiment. By analyzing the data from the Minnesota Challenge, we hope to answer the following questions:

(1) How much was the game teaching? For example, if pass rates for fastforward Challenges were the same if received in the first few levels compared to the standard point in the progression, this would be evidence the game was not improving student ability. Much like a pretest, the overall pass rate early on could also give us an indication of how much the underlying player population knew.
(2) What effect do early mastery tests have on player mastery rates and engagement (measured in total time played), among players who pass them? If the boredom hypothesis is correct and a significant problem, then early mastery tests should increase both player engagement and mastery rate by preventing boredom in players who pass the tests.

(3) What effect do early mastery tests have on player engagement among players who fail them? According to the theory of flow [Csikszentmihalyi 1990], boredom and frustration are at odds: receiving and failing a difficult test could discourage players, possibly overpowering any positive effects of fastforwarding. In the best case – players are unaffected by failing a test – we would even have supporting evidence for using these types of difficult mastery tests as pretests in a campaign-game environment. This could potentially overcome one of the primary difficulties with game-based research: the difficulty in measuring learning gains due to lack of pretest-posttest data.

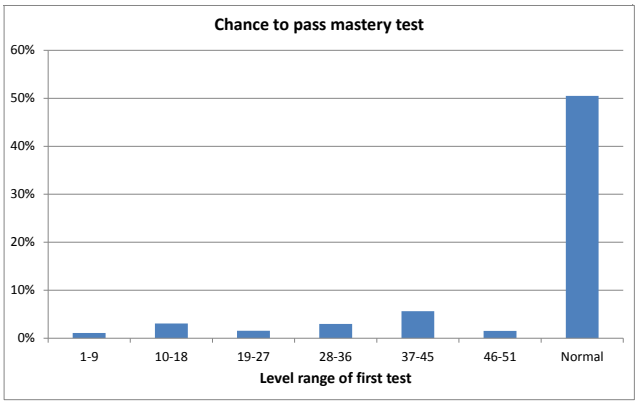
We hypothesized that effects on mastery rate and student engagement might change depending on when the fastforward tests showed up, with the strongest effects early on. In the standard Minnesota progression, when no random early tests were given, players had to play a minimum of 52 levels to reach the three mastery tests. Though we note that due to adaptivity that would insert extra levels, students often played considerably more than the minimum number of levels: on average students receiving the mastery tests at the normal point had to play 99 levels to do so.

We then asked what effect triggering the mastery test on the n'th level would have. Given the low probability of receiving the test at any particular time, we binned the early levels into groups of 9. For any given bin, such as 10-18, we collected all the players who played at least 18 levels, then divided them into two groups: those who received no tests after the first 18 levels, and those who received no tests after the first 9 levels and at least one test after levels 10-18 (further subdivided into those who passed the tests and those who didn't). These groups correspond to the control, the highly adept who might have been bored, and those who might have been harmed by being given the test too early. The last bin, 45-51, is shortened to avoid considering players who played perfectly and received a normal test after level 52.

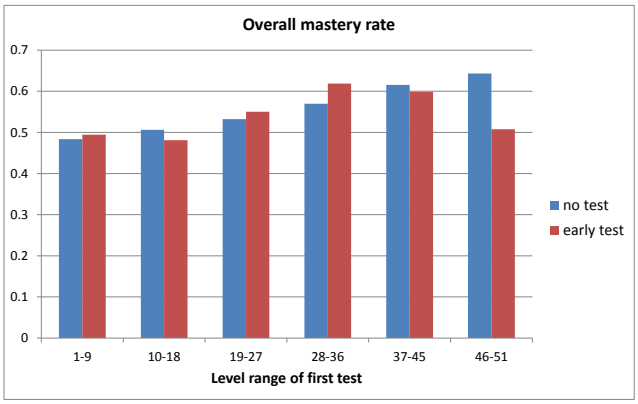
The results can be seen in Figure 5. By checking how likely players were to pass tests when given in different bins, we see from Figure 5(a) that players were extremely unlikely to pass the tests when given very early on, especially when compared to pass rates on tests reached normally through the progression. We found this quite surprising. Note that while it was possible for students to receive a test by level 52, very few students actually did so because the adaptive level progression would give them extra practice when they had trouble solving levels. Thus, nearly every student who received an early test, even in the 45-51 bin, had not yet reached levels with key concepts used in the mastery test. This fact, combined with the low overall pass rates, implies both that players did not already know everything the game could teach, and also that we would expect any potential gains from fastforwarding players to be minimal because so few players pass.

This turns out to be the case. By checking eventual mastery rates of players given tests in the various bins, we see from Figure 5(b) that the eventual mastery rates of those given or not given the test in any particular bin are very similar. Note that the mastery rates continue to increase from bin to bin because we only study players who make it to the end of each bin, so that at later bins players who quit before achieving mastery are removed. χ^2 tests confirm that giving early tests has no effect on eventual mastery rates, as seen in the upper part of Table V. We conclude that our fastforward tests did not achieve the intended purpose of increasing mastery rates by allowing players to skip boring content.

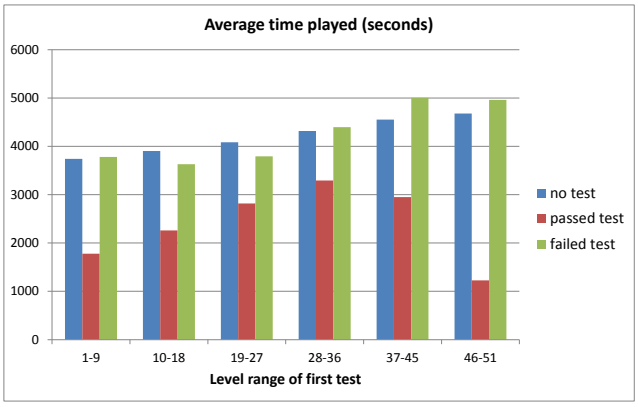
We also asked whether there would be negative effects if students were given very difficult levels early on. This can be analyzed by comparing total time played between students who were not given tests and students who were, as seen Figure 5(c). We caution that only a handful of students were able to pass the fastforward tests. These



(a) Pass rates



(b) Overall mastery



(c) Engagement

Fig. 5: Results of randomly giving out mastery tests at the beginning of the game, where “no test” means that students did not receive an early test level by the end of that level block (thought they might receive one later). Pass rates are very low compared to pass rates when reaching the tests through the normal progression; early tests appear to have little effect on eventual mastery rates; and failing tests has little effect on player engagement.

Early test bin	χ^2 statistic	p-value ($\alpha = 0.004$)
1-9	$\chi^2(1, N = 6804) = .159$	.365
10-18	$\chi^2(1, N = 6168) = .706$	.218
19-27	$\chi^2(1, N = 5582) = .330$	.305
28-35	$\chi^2(1, N = 4952) = 2.194$	.078
36-45	$\chi^2(1, N = 4371) = .375$	.375
46-51	$\chi^2(1, N = 4059) = 5.105$	.018

Early test bin	Wilcoxon Z statistic	p-value ($\alpha = 0.004$)
1-9	$Z = -.125$	.901
10-18	$Z = -2.373$	.018
19-27	$Z = -.593$	.553
28-35	$Z = -.215$	.830
36-45	$Z = -1.213$	.225
46-51	$Z = -.189$	.850

Table V: Post-hoc analyses run to understand the effect of giving early mastery tests; with 12 comparisons, the Bonferroni correction gives us critical $\alpha = 0.004$. The top table compares eventual mastery rates of players given early tests against those not given tests at those times. The bottom table compares time played by students who are given early tests and fail them against those not given tests; the Wilcoxon rank sum test was used due to non-normality of the data. No results are significant.

students tended to play fewer levels, but there is insufficient data to draw firm conclusions; we can only speculate that they may have simply had less content because they skipped many levels, or perhaps the tests merely identified students who would not have played long anyways because they knew the content the game was meant to teach.

Interestingly enough, we discovered no statistically significant effect on engagement when students failed the early tests, as seen in the bottom half of Table V. If true, this result suggests that giving pretest levels very early may be a viable way to estimate student learning without causing negative consequences for player engagement; it could also mean that games or adaptivity schemes will not cause too much harm to the student population when students are occasionally given extremely difficult problems *if they can submit an incorrect answer without penalty*.

Regardless of the specific lessons learned from our fastforward mastery test, a more general lesson learned from the Minnesota Algebra Challenge is that campaigns can be successfully used to run randomized controlled experiments. Without any competitive incentives, we were able to add extra assessment levels that could have drastic impacts on players' progress through the game, and received no complaints about this behavior. We caution, however, that players were all basically performing the same task and would have been very unlikely to notice or discuss differences in the levels they saw. Experiments with drastically different conditions would be more difficult to run.

6. CONCLUSION

In this paper, we discuss a new method for delivering educational technology: the *campaign*. Campaigns are focused, relatively short events designed to promote and encourage the use of educational technology across a wide variety of students, and are quite different from other popular methods of delivering such technology or content, such as downloadable software, free online websites, or MOOCs. In light of the high rate of mastery (96%) among students in our campaigns who played the 1.5 hours we re-

quested, they seem deserving of research as educational tools. Furthermore, while we have framed campaigns as being useful for delivering educational technology, it may be possible to use them with other types of content or software. This would require careful incentive design for users to participate, which we leave to future work.

To better understand the properties of campaigns, we present a case study using three campaigns we conducted with the educational game DragonBox Adaptive: the Washington, Norway, and Minnesota Algebra Challenges. We detail the costs and logistics of running campaigns, and describe basic information about student participation and achievement in the form of time played and ability to achieve mastery of the content. To help others planning on running campaigns, we share several of our observations. First, players play our game orders of magnitude longer than they have played other educational games offered on free websites. Second, collaborative incentives and rewards to classes for achieving mastery of the content may not have much effect on how long students play, but competitive incentives and rewards to students for finishing levels may have large effects while also leading to undesirable "gaming" behavior. Third, running experiments using campaign data can be challenging due to the difficulty of randomizing within schools or teachers, and campaigns may have differential effects depending on student gender and socioeconomic status. Finally, giving students "pre-test" levels they are overwhelmingly likely to fail does not necessarily cause frustration or other negative effects. Each of these findings requires further study, and suggests other interesting lines of research, but taken together are a promising first step in understanding how campaigns can be used to achieve learning at scale.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the Office of Naval Research grant N00014-12-C-0158, the Bill and Melinda Gates Foundation grant OPP1031488, the Hewlett Foundation grant 2012-8161, Adobe, and Microsoft.

REFERENCES

Erik Andersen, Yun-En Liu, Richard Snider, Roy Szeto, and Zoran Popović. 2011. Placing a Value on Aesthetics in Online Casual Games. In *CHI '11: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA.

John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* 4, 2 (1995), 167–207.

Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. 2004. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems*. Springer, 531–540.

Lisa Barrow, Lisa Markman, and Cecilia E Rouse. 2008. *Technology's edge: The educational benefits of computer-aided instruction*. Technical Report. National Bureau of Economic Research.

Larry Berger and David Stevenson. 2007. K-12 entrepreneurship: slow entry, distant exit. *American Enterprise Institute Conference on the Future of Educational Entrepreneurship* (2007).

Lori Breslow, David E. Pritchard, Jennifer DeBoer, Glenda S. Stump, Andrew D. Ho, and Daniel T. Seaton. 2013. Studying Learning in the Worldwide Classroom: Research into edXs First MOOC. *Research and Practice in Assessment* 8 (2013), 13–25.

Center For Game Science. 2014. DragonBox Adaptive. (May 2014). http://centerforgamescience.org/portfolio/dragonbox/

Aaron Chatterji and Benjamin Jones. 2012. Harnessing Technology to Improve K-12 Education. *Policy Brief* 5 (2012).

A. Corbett, K. R. Koedinger, and J. R Anderson. 1997. Intelligent Tutoring Systems. In *Handbook of Human-Computer Interaction, Second Edition*, M. Helander, T. K. Landauer, and P. Prahu (Eds.). Elsevier Science, Amsterdam.

Council of Economic Advisers. 2011. Unleashing the Potential of Educational Technology. (2011).

Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. Harper & Row Publishers, Inc., New York, NY, USA. 49 pages.

Dominique Foray. 2011. Educational Innovation: An Economists Perspective. *Rigour and Relevance in Educational Research* (2011), 35.

Roland G Fryer. 2011. Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics* 126, 4 (2011), 1755–1798.

James Paul Gee. 2005. Learning by Design: Good Video Games as Learning Machines. *E-Learning and Digital Media* 1, 1 (2005), 5–16.

James P. Gee. 2008. *What Video Games Have to Teach Us About Learning and Literacy*. St. Martin's Press. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/1403984530

Philip J. Guo and Katharina Reinecke. 2014. Demographic Differences in How Students Navigate Through MOOCs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference (L@S '14)*. ACM, New York, NY, USA, 21–30. DOI:http://dx.doi.org/10.1145/2556325.2566247

Grant D. Jacobsen and Kathryn H. Jacobsen. 2011. Healt awareness campaigns and diagnosis rates: Evidence from National Breast Cancer Awareness Month. *Journal of Health Economics* 30, 1 (2011), 55 – 61. DOI:http://dx.doi.org/10.1016/j.jhealeco.2010.11.005

Kenneth R. Koedinger, John R. Anderson, William H. Hadley, and Mary A. Mark. 1997. Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education* 8 (1997), 30–43.

Ron Kohavi, Randal M. Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 959–967. DOI:http://dx.doi.org/10.1145/1281192.1281295

Tamar Lewin. 2012. Education site expands slate of universities and courses. *The New York Times* (19 September 2012).

Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popović. 2014. Towards automatic experimentation of educational knowledge. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3349–3358.

Derek Lomas, Kishan Patel, Jodi L. Forlizzi, and Kenneth R. Koedinger. 2013. Optimizing Challenge in an Educational Game Using Large-scale Design Experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 89–98. DOI:http://dx.doi.org/10.1145/2470654.2470668

Merrilea J. Mayo. 2009. Video Games: A Route to Large-Scale STEM Education? *Science* 323 (2009), 79–82.

Robert Murphy, Larry Gallagher, Andrew E. Krumm, Jessica Mislevy, and Amy Hafter. 2014. *Research on the Use of Khan Academy in Schools: Implementation Report*. Implementation report. SRI International.

Harold F. O'Neil, Richard Wainess, and Eva L. Baker. 2005. Classification of learning outcomes: evidence from the computer games literature. *The Curriculum Journal* 16, 4 (2005), 455–474.

E. O'Rourke, E. Butler, Y. Liu, C. Ballweber, and Z. Popović. 2013. The Effects of Age on Player Behavior in Educational Games. In *Foundations of Digital Games (FDG '13)*.

Eleanor O'Rourke, Kyla Haimovitz, Christy Ballweber, Carol S. Dweck, and Zoran Popović. 2014. Brain Points: A Growth Mindset Incentive Structure Boosts Persistence in an Educational Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*.

Alexandros Paramythis and Susanne Loidl-Reisinger. 2004. Adaptive Learning Environments and eLearning Standards. *Electronic Journal of eLearning* 2 (2004), 181–194.

A.M. Pullybank, N. Dixon, and A.R. Dixon. 2002. The impact of bowel cancer awareness week. *Colorectal Disease* 4, 6 (2002), 483–485.

S.W. Raudenbush. 1988. Educational Applications of Hierarchical Linear Models: A Review. *Journal of Educational Statistics* 13, 2 (1988), 85–116.

Katharina Reinecke and Krzysztof Z. Gajos. 2014. Quantifying Visual Preferences Around the World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 11–20. DOI:http://dx.doi.org/10.1145/2556288.2557052

Cecilia Elena Rouse and Alan B Krueger. 2004. Putting computerized instruction to the test: a randomized evaluation of a scientifically based reading program. *Economics of Education Review* 23, 4 (2004), 323–338.

Valerie Shute and Brendon Towle. 2003. Adaptive E-Learning. *Educational Psychologist* 38, 2 (2003), 105–114.

E. A. Skinner and M. J. Belmont. 1993. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology* 85, 4 (1993), 571––581.

John C Stamper, Derek Lomas, Dixie Ching, Steven Ritter, Kenneth R Koedinger, and Jonathan Steinhart. 2012. The Rise of the Super Experiment.. In *EDM*. 196–200.

Kurt VanLehn. 2006. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education* 16 (2006), 227–265.

Kurt Vanlehn, Collin Lynch, Kay Schulze, Joel A. Shapiro, Robert Shelby, Linwood Taylor, Don Treacy, Anders Weinstein, and Mary Wintersgill. 2005. The Andes physics tutoring system: five years of evaluations. In *In Proceedings of the 12th international conference on Artificial Intelligence in Education*. IOS Press, 678–685.

K. Wauters, P. Desmet, and W. Van den Noortgate. 2010. Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning* 26, 6 (2010), 549 – 562.

We Want To Know. 2014. DragonBox. (May 2014). http://www.dragonboxapp.com/

H. Woltman, A. Feldstain, J.C. MacKay, and M. Rocchi. 2012. An Introduction to Hierarchical Linear Modeling. *Tutorials in Quantitative Methods for Psychology* 8, 1 (2012), 52–69.