Evaluating the Effectiveness of Animated Cartoons in an Intelligent Math Tutoring System Using Educational Data Mining

Giora Alexandron
Weizmann Institute of Science
Israel

Gal Keinan
The Center for Educational Technology
Israel

Ben Levy
The Center for Educational Technology
Israel

Sara Hershkovitz
The Center for Educational Technology
Israel

Abstract: We studied the effectiveness of animated cartoons in an Intelligent Tutoring System (ITS) that teaches Fractions for 4th grade. By examining viewing patterns, as well as the relationship between viewing and performance on related tasks, we sought to better understand how young students learn mathematical concepts with cartoons. The analysis serves two purposes: The first is to optimize the pedagogy of the specific ITS. The second is to derive general pedagogical principles that will guide the design of future ITSs. This study is part of a larger effort aimed at enhancing our understanding of how to make better digital learning environments using educational data mining. The study uses data from 650 4th grade students who used the ITS for 4-8 weeks. The analysis is based on combining high-resolution clickstream data with meta-data that describe the content and its structure. To measure the relationship between viewing and performance, we used a between-group quasi-experiment design as well as regression analysis.

Introduction

In recent years there has been a technological shift in education towards extensive use of digital learning environments, such as Massive Open Online Courses (MOOCs) and Intelligent Tutoring Systems (ITSs), which can provide a more personalized learning experience. Together with increasing its use, it is important to enhance our understanding of how to design *effective* digital learning experiences. It is hoped that the fine-grained data that such digital platforms collect can help to close this gap, by providing insights into learners' behavior, and by providing a means to evaluate and improve the specific context – course and platform – that generated the data.

We describe a step in this direction within the domain of K6 Math Education by presenting a study that applies educational data mining techniques to a new Intelligent Tutoring System that teaches fractions in 4th grade. Specifically, the study focuses on the use of the instructional animated videos (cartoons [1]) that are included in the digital course. This analysis serves two pedagogical purposes. The first evaluates the effectiveness of the cartoons in this specific course. The second derives more general insights that can guide future instructional design of other digital courses.

The analysis uses clickstream data that contain, among other things, videos, navigation, and submission events, which the ITS records after each step. Using these fine-grained data, we can obtain direct measurements of how students use the animated cartoons, and link them to students' performance on tasks related to each cartoon, namely, the essential concepts and skills that this cartoon introduces. This educational data mining methodology for evaluating the pedagogical value of videos resembles the one that was used by Chen et al. (2016) for evaluating different video shooting styles in a MOOC. In both cases, the analysis also relies on close collaboration between the content team, which adds their expertise to the domain and to the discipline-based research, along with the data science team, which adds expertise in data mining and statistical modelling. Our modelling of the in-video behavior is quite similar to the ones described in (Kim et al., 2014, 2016), who focused on in-video activity. The experimental set-up is based on a quasi-experiment involving between-groups design and regression analysis, similar to the one in (Roll, Baker, Aleven, & Koedinger, 2014).

[1] The course used animated cartoons, which are a specific genre of videos. In general, we used the term 'video' to refer to the media type, and 'cartoon' to the specific application in this ITS. However, since the methodology that we present is applicable to any type of video, we sometime used the more general term.

The findings, reported in Section 5, show that watching the cartoons was associated with a mean increase of ~8-9% in the performance of the related activities. This increase was consistent in both models. In addition, the analysis revealed some interesting use patterns. For example, a second-by-second analysis revealed that dropout tended to increase after about 90 seconds of video, and that male students showed higher dropout rates.

We believe that this paper contributes in several directions. On the pedagogical side, it provides evidence that short cartoons can be an effective means to explain mathematical concepts in K6 self-learning environments, and it reveals some interesting use patterns. On the methodological side, it suggests a methodology for evaluating the contribution of videos, and demonstrates how educational data mining techniques can be applied in commercial settings to evaluate and improve the learning experience. To the best of our knowledge, this paper is the first to report on such an analysis of clickstream data to evaluate the effectiveness of videos in K12 ITSs.

The remainder of the paper is organized as follows. Section 2 surveys related literature. Section 3 describes the experimental set-up. Section 4 describes the data. The results are presented in Section 5, followed by the Discussion and Conclusions in Section 6.

Related Work

Animated cartoons or videos are widely used in K12 STEM education to teach scientific concepts (Dalacosta, Kamariotaki-paparrigopoulou, Palyvos, & Spyrellis, 2009). For example, BrainPOP (https://www.brainpop.com/) is a popular educational website with over 1000 short animated educational movies for K12. Understanding the effectiveness of such cartoons as a means for teaching scientific concepts is important for instructional and media designers who develop and use such cartoons. Barak, Ashkar, & Dori (2011) reported that BrainPOP videos promoted students' understanding of scientific concepts, and increased their motivation to learn science. The experimental set-up was based on a year-long, between-group experiment with a treatment group (class) in which the teacher used BrainPOP videos to explain scientific concepts, and a control group (class) who learned the conventional way. A positive effect of using animated cartoons in K12 science teaching was also reported by (Dalacosta et al., 2009), based on a one-time controlled experiment with a control group who received a conventional explanation, and a treatment group who received explanations from animated cartoons.

However, we studied the issue in a different pedagogical context and with very different methodological means. Regarding the context, students are in class but learn from an Intelligent Tutoring System – a self-learning environment in which students are free to progress at their own pace. The methodology is based on mining fine-grained clickstream data. Analyzing such data enables us to learn how and when students use the cartoons, whether they return to specific points, where they dropout, and more. To date, analyzing high-resolution data to study how students learn from videos in self-learning environments has been conducted in Massive Open Online Courses (MOOCs). For example, Kim et al., (2016) and Guo, Kim, & Rubin (2014) analyzed in-video behavior in order to derive design guidelines, with engagement as the main metric. Chen et al. (2016) measured the pedagogical efficacy of different video styles by intersecting in-video behavior with performance on related tasks. Machine-learning methodologies for measuring the effectiveness of videos (and other instructional resources) were suggested in (Alexandron & Pritchard, 2015) and (Machardy & Pardos, 2015). In a learning context that is more similar to ours (Intelligent Tutoring system for K6 science education), a methodology for evaluating a different user's behavior – seeking help – was proposed in (Roll et al., 2014).

Outside the educational context, using data mining, direct measurement and quasi-experimental methods to study video-related user behavior such as the effectiveness of video ads (Krishnan & Sitaraman, 2013), and more generally, for Causal Discovery (Oktay, Taylor, & Jensen, n.d.), are much more common. We found some of the quasi-experimental methodologies applied in social media to be less appropriate for our learning scenario because of its fragmented (many sub-topics) and dynamic nature. Thus, we combine such methods with the methodology of (Roll et al., 2014)

The Course and the Experimental Set-up

Context

The experimental set-up is based on a pilot digital course in which 4th grade students learned the Fractions portion of the 4th grade Math curriculum from a new self-paced Intelligent Tutoring System (ITS). This pilot course replaces the conventional way in which the subject is taught, which relies mainly on frontal teaching. The pilot included ~650 learners from ~20 classes of varied demographics, who participated in two 45-minute lessons a week, during regular Math lessons, for 4-8 weeks. The pilot took place during the last two months of the 2016-2017 school year. The teachers were present in class during all the lessons, and aided the students when needed. They received training prior to the pilot, and support during it, from the online instruction field support team.

The first goal of the pilot was to evaluate whether students who learn with ITS reach at least the same level of mastery in the topic as students who learn in the standard fashion. Namely, to verify that the new pedagogy does not lead to degradation in students' learning. (Obviously, this is the minimal requirement; the goal is to suggest a better pedagogy that improves the current level.) This issue was analyzed separately and is beyond the scope of the current paper.

The second goal was to collect empirical, fine-grained data on students' behavior within the tutor, and use it in two ways. One is to develop learning analytics pipelines for improving the content and the platform. The second is to design data-driven personalization mechanisms for the next phase of the product. The analysis that is the basis for this paper was conducted as part of this dual effort. More details on the data and their collection are presented in the next section (Data and Methods).

Course structure and content

The course covers the 4th-grade portion of the Fractions syllabus of the Israeli Math Curriculum for elementary schools (the topic is further studied in 5th and 6th grades). A major factor for choosing 'Fractions for 4th grade' for the pilot was that it is the first introduction to this topic within the curriculum; thus, it provides a 'cleaner' set-up that is less affected by previous learning.

The course implements an active learning pedagogy. It includes 112 self-contained learning units, each composed of a series of activities centered around a specific sub-subject (e.g., "multiple representations of the same fraction"). Some of the activities are accompanied by 'labs', which are interactive learning aids that enable the students to explore the mathematical properties that are being learned, raise hypotheses, and test them. The learning units are arranged in a linear order. Students can progress in a non-linear order, but besides a few exceptions, the students followed the linear order.

The cartoons

Some of the learning units begin with short (1-3 minute) animated cartoons, which introduce new topics, present worked examples and different representations, etc. They were developed and produced in collaboration with the content developers and video experts. Overall, there are 18 cartoons, located at the beginning of 20 learning units. The length of the cartoons varies from 1 to 3 minutes, with most of them from 1.5 to 2.5 minutes.

All the cartoons use the same animated characters: a child who deals with Math challenges, a giraffe that explains how to solve them, and a lazy tiger that provides some comic relief. The cartoons follow a similar plot: an opening of ~10 seconds; a ~30-second introduction to the topic, presented by the child, through a challenging problem that needs to be solved; an explanation, presented as a problem-solving dialog between the child and the giraffe; and a closing event. A link to a representative cartoon is provided below[2]. Figure 1 shows two screenshots from a representative cartoon. On the left, we see a dialog between the child and the giraffe, and on the right, a detailed explanation given to the child by the giraffe.

Figure 1: Screenshots of the dialog (left) and the explanation (right)

[2] https://www.youtube.com/watch?v=mVsDx1pyV7U&feature=youtu.be

Success on the first attempt

We used a dichotomous model in which the performance of student 's' on question 'q' was considered as a success on the first attempt (this holds for the analysis; students can have multiple attempts). Using 'success on the first attempt' as a proxy for performance is a common practice in online learning environments that allow multiple attempts (Bergner, Colvin, & Pritchard, 2015). A question that the student skipped is counted as incorrect (when the student does not return to it afterwards). A question that the student did not see is not counted [3](did not see is operationalized as did not enter the page containing the question).

Local measure of learning

In order to operationalize the notion of learning in a way that can be quantified and associated with watching the cartoon, we examined students' performance and the sequence of questions that followed the cartoon. We used two measures: The first is based on the first question after the cartoon. This is similar to the 'local measure of learning' used in (Roll et al., 2014) for evaluating the benefit of seeking help (we adopted the term 'local measure of learning'); the second measure that we used is the performance of the entire sequence that follows the cartoon (typically around 5 questions).

The rationale for using the first question is that it is closely related in terms of content, and that the proximity in time between watching and answering reduces the risk of a confounding effect. The rationale for using performance for the entire sequence is to evaluate whether students who did not watch the cartoon can 'catch up' by just answering more questions (if that would be the case, the cartoon might be redundant).

Data and Methods

Below we describe the data collection mechanisms, the raw and meta data, and the quasi-experiment analysis methods that we used.

Data

Raw and Meta-Data

The interaction of the learners with the learning environment is captured via Experience xAPI [4], which logs the users' clickstream activity. Specifically, in this study we relied on video (*play, pause, seek, stop*), navigation (browsing pages), and question attempting events. The structure of the course and the relationship between its components (e.g., which question resides on which page) are captured in the course structure files that are generated automatically by the Learning Management System (LMS). (In addition, the pedagogical team developed a taxonomy that models the domain knowledge that students need to master, along with pedagogical aspects such as interaction type, and they mapped the questions into it. Later this taxonomy will be extended with empirical meta-data such as the level of difficulty. However, this taxonomy was not used in the analysis presented here.) Overall, the 650 students in the course generated about 2 million events. Of these, about half a million are question responses, and about 100 thousand are video-related events.

Video Viewing Sessions

For each student, we collected the viewing sessions. A session is defined as a single, continuous watching interval. It is represented as a tuple <*student, video, time, start_location, end_location*>. A *start_location* represents the location (in seconds) within the video in which the session begins, for example, when the user plays the video from its beginning, *start_location* =0. *End_location* represents the location within the video in which the current watch session ends. *Time* represents the global time when the session starts. *Student* and *video* are unique identifiers in this video, and for the student, within the system.

The sessions are collected by mining the clickstream data, in a fashion similar to the one described in (Guo et al. 2014). Each *video_play* event opens a new session, and the session is 'closed' when one of the following happens: i) the user emits a *video_pause* event; ii) the video finishes (a *video_stop* event is emitted the system); iii) the user triggers an event not related to the video, such as navigating to the next page; iv) the user drags the video slider or clicks on another location in the video progress bar (here, the current interval is closed, and a new one is opened). We noted that a small number of video events included corrupted data, and some of the events were clearly

[3] This is basically a finer treatment of missing data that is finer than is interpreting missing data as either *Incorrect* or *Missing at Random* (Bergner et al., 2015)

[4] https://www.adlnet.gov/xAPI

lost. This is typical of clickstream logs, and especially regarding parts of it that rely on information from external sources (here, *YouTube*). To address this issue, we deleted 'broken' sessions.

Next, we built, per student *s* and video *v*, the range/s of seconds within *v* that *s* watched (this can be a fragmented collection of intervals). This was achieved by performing a *union* operation on the set of *start_locations* and *end_locations* of the sessions of *s* on *v*. This results in a set of disjoint time intervals. Eventually, the index that we used is the *percentage of v that s watched*, denoted as <*s, v, perc_s_v*>.

Quasi-Experiment Analysis

To differentiate between watching the cartoons and performance, we used two Quasi-Experiment methods. The data set that was used for this analysis includes the students who completed at least the first of two parts of the course (67 out of the 112 learning units). This is to ensure that the research population is relatively stable and to omit potential bias due to inclusion of data from students who might have low motivation.

Between-groups comparison

The first model that we used is a between-groups comparison. For each learning unit that contains a cartoon, we compared the performance of those students who watched the cartoon ('the treatment group') to the performance of those students who did not watch it ('the control group'). Performance was measured as the proportion of correct performances on the first attempt for the sequence of questions that follow the cartoon. Student *s* is considered as 'having watched' cartoon *v* if *s* watched more than 75% of *v*, and *s* is considered as 'did not watch' *v* if *s* watched less than 25% of *v* (watching sessions in the 25-75% range were omitted).

Regression Analysis

We used a methodology that modifies the one used by Roll et al. (Roll et al., 2014) to evaluate help-seeking behaviors. The measure of learning that we used is *correct on the first attempt* for the first question after the cartoon, similar to their notion of 'local measure of learning'. Our model attempts to distinguish between learning and watching, while adjusting for student and question parameters. To do so, we fitted the following logistic regression model:

$$Log\ Odds\ [CFA(Si,Qj)] = B0 + B1 * Si\_ability + B2 * Qj\_difficulty + B3 * Si\_watched\_Vj$$

This function models the *Log Odds* of Student *Si* as being correct on the first attempt with *Qj* (the first question after cartoon *Vj*), as a linear combination of *Si*'s ability, *Qj*'s difficulty, and whether *Si* watched *Vj*.

B0 is the intercept. *B1* adjusts to the student's ability, operationalized as the proportion of correct solutions on the first attempt for questions that the student encountered so far (previous units in the course). *B2* adjusts for the difficulty level of the question, which is operationalized as the proportion of correct solutions on the first attempt for this question. Last, *B3* evaluates the effect of seeing the cartoon. It is a Boolean variable that receives True iff *Si* watched more than 75% of *Vj*. $B3 > 0$ means that seeing the cartoon is associated with an increased likelihood of solving the question correctly.

The main difference between our regression model and the regression model used in Roll et al. is that we *parameterized* the students and activities (according to their ability and difficulty, respectively), whereas Roll et al. used variables per *student* and *skill*. This explicitly controls for between-student variations, and as they noted, it does not capture *within-student* variations. Importantly, our modeling controls for between-student variations using a parameter [5], and it can capture within-student changes in the behavior during the course[6].

Overall, the data set for the regression model contains 4984 rows of 345 students attempting to solve 17 questions (we omitted the second occurrence of the 2 cartoons that were used twice, and the cartoon from the first unit, since for this cartoon no previous ability was found to use it). The logistic regression was trained using the standard *glm* function in R. The ability and the difficulty parameters were transformed from a 0-1 scale (proportion correct on the first attempt) to a standard z-score.

---

[5] In practice, we also fitted a regression model that fits a parameter per student and per question. The results (with respect to the coefficient of the *watching* covariate) were almost identical to the ones reported in Subsection 5.2.

[6] Parameterizing ability as performance from the beginning of the course smoothens local changes. However, we fitted also a different model having a local performance parameter (operationalized as performance in the previous unit), and obtained results (with respect to the coefficient of the *watching* covariate) similar to the ones reported in Subsection 5.2.

Results

This section is organized as follows: First, we present descriptive statistics regarding the level of viewing and dropout, and their intersection with factors such as overall performance, gender, and school/class affiliation. Second, we present the results regarding the relationship between watching the cartoons and the performance on related items.

Amount of Use

Amount of view, per cartoon

Figure 2 shows the percentage of students that opened, reached the middle of, and completed, each of the cartoons. For each cartoon, 100% is the number of students who accessed the unit in which the cartoon resides (upper gray line, with the numbers that constitute 100%). For example, the first cartoon was opened by 80% of the 652 students who accessed the unit in which it resides (the first unit), 60% of the students viewed at least half of it, and 55% viewed it completely (viewing more than 90% of the video was considered as 'viewed completely', including those students who closed the cartoon in the last few seconds, which returns to the non-mathematical part of the cartoon).

As can be seen, the patterns are relatively stable over time, with a slight downward trend. Two exceptions are videos 5 and 8, for which there was a sharp drop in the number of students who viewed at least 50% of the video. After this was checked with the content developers, it turned out that these cartoons were actually the second use of a cartoon that was used in a previous unit. Students noticed this after a few seconds and skipped the remaining part (which indicates that the watching was learning oriented; we referred to that in the Discussion). We left them in the graph to demonstrate some of the anomalies that can be identified when visually inspecting dropout graphs. However, these two duplicates, and the questions (learning sequence) that are associated with them, were removed from the rest of the analysis.

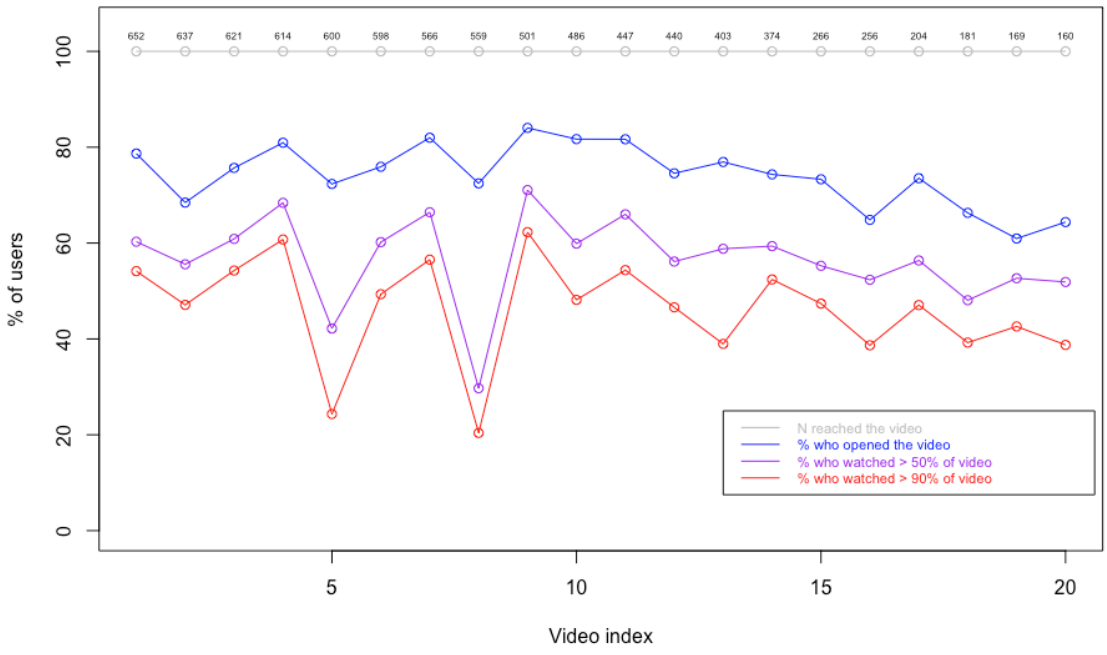

Figure 2: Amount of view per cartoon

Next, we were interested in determining the distribution of the number of views among the students. For example, we wanted to know whether the ~20% of students that did not open each video (the gap between the '% opened' and the 100% in Figure 2) are the 'same' students. Figure 3 presents the distribution of the percentage of videos opened and watched by more than half of the students. As can be seen, more than 95% of the students opened 80-100% of the videos to which they had access. Only 1% of the students did not open any of the videos. This clearly indicates that the ~20% of the students that did not open each video were NOT the 'same' students, but

rather, were different ones (the 'same' students not opening many videos would result in a bimodal distribution with a high bar on the left side).

In-video dropout

Continuing to study students' dropout patterns, we examined the dropout over time within the video. To do so, we assembled the sessions watched, and compared, in 10-second intervals, the percentage of students who dropped at this time interval with the percentage of those students who started it. To reduce noise owing to the varying length of the videos, we considered in each interval only those sessions on videos that are longer than the end of the interval, with some 'extra' to reduce end-of-video dropout noise (for example, the percentage of dropouts in the '90-100 second interval' was computed from videos that are longer than 110 seconds). This is presented in Figure 4. The dotted line denotes the exact values, and the gray line is a smoothing curve. It shows that in the first 10 seconds the dropout is relatively high (>8%), then it is goes down, and starts to climb back after ~90 seconds.

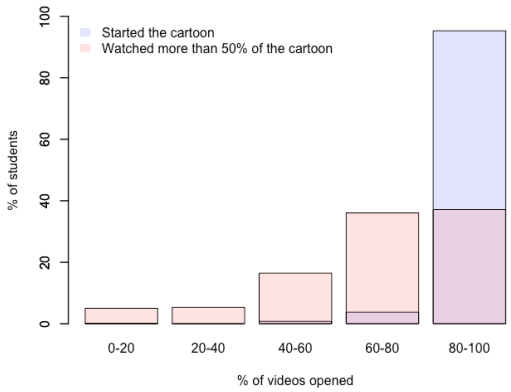

Figure 3: Distribution of opened/watched videos

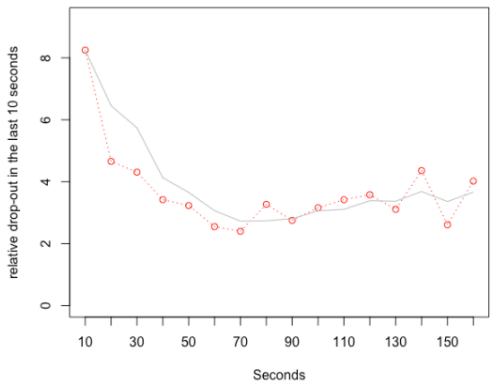

Figure 4: In-video dropout

School/Class factors

One of the issues that the instructional designers and the field instructors were interested to clarify is whether school and class factors play an important role in the use of the cartoons. Figure 5 shows, per school (left chart) and the classes within it (right chart), the distribution of the percentage of videos that each student watched ('watched' is operationalized as having viewed more than half of the cartoon; classes are denoted by the same color as their school). For example, school *J* (in gray) has 40 students (*N*, above the boxplot), and a median of 60%, meaning that half of the students watched more than 60% of the videos to which they had access. School *J* includes two classes – *J1* and *J2* (also in gray), of 20 students each. When the boxplot of *J1* and *J2* is observed, a significant difference between the classes is evident. Whereas the median of *J1* approaches the median of the rest of the classes, the median of *J2* is significantly lower. *J2* is a religious school with single-gender classes. The class with the low viewing rates is the class consisting of male students. According to the instructor, who worked with both classes, the students in the male class said that they preferred to progress and did not want to 'waste time' by watching the videos.

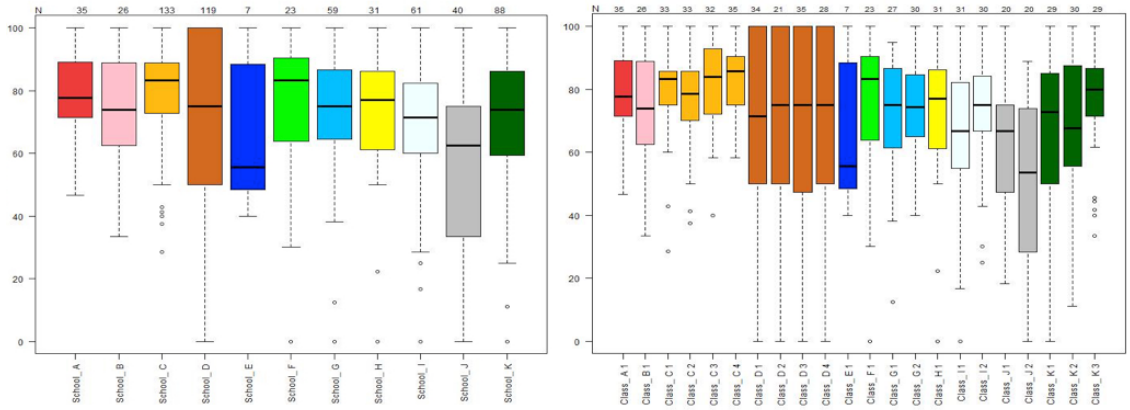

Figure 5: Viewing per school (left) and class (right)

Another school with very low use is the 'blue' school. In this school, we had one class of 7 students with special needs. We intend to study the use patterns within this population in more depth in the future.

Gender Differences

Following the findings that among the single-gender classes (*J1* and *J2* in Figure 5), the number of views in the male class was significantly lower, we wanted to determine whether this pattern is repeated in the mixed classes namely, if male students tended to watch fewer cartoons. We found that, on average, male students tended to watch 71% of each cartoon, whereas for females, the average was 77%. A two-tailed t-test confirmed that the distribution of the length of the sessions by male and female students is not equal, with a *p-value <0.001*.

The differences in the extent of watching in the two groups is presented in the two figures below. Figure 6 presents the accumulated distribution function of the length of the sessions watched by male and female students. For example, 29% of the sessions in the male group ended before they reached half of the video, whereas in the female group, only 24% of the sessions ended before that point.

Figure 7 shows the results aggregated per student. It presents the distribution of the proportion of cartoons opened, watched half way, and watched entirely (operationalized as >90%, to remove the noise of exiting at the last second), among the cartoons to which each student had access. For example, among the female students the median for the half-way point is 0.77. This means that 50% of the female students reached the midpoint of 77% of the cartoons to which they had access. Among the male students, the median for the midpoint was lower, 73%. As can be seen, the difference between the groups at the end of the cartoons is considerably larger (the medians are 0.67 and 0.57 for females and males, respectively).

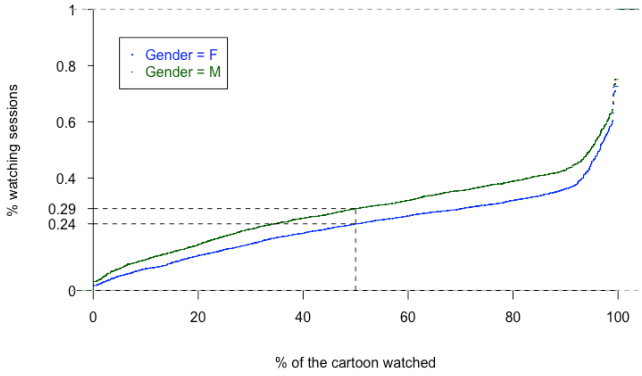

Figure 6: Distribution of session length

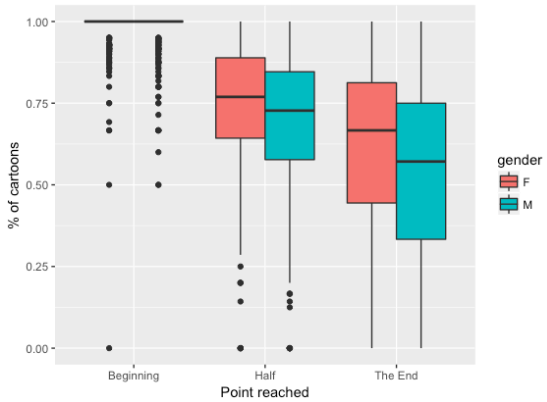

Figure 7: Distribution of views per student

Correlation with the overall performance in the course

In order to determine whether viewing the cartoons is a pattern associated with students' performance, we examined the correlation between watching the cartoons and the overall success. Here, cartoon v was counted as 'watched' by student s if s viewed more than half of v, and success was computed as the proportion that was correct on the first attempt in the entire course. This yielded a weak positive correlation of 0.15.

The relationship between watching the videos and performance in related tasks

Between-Groups comparison

As described in Subsection 4.2, this model compares, per cartoon, the performance of the group of students who watched the cartoon to the performance of the ones who did not watch it. This yielded 18 pairs of mean values. A paired t-test confirmed that the performance of the group who watched the cartoons is higher. The mean difference between the groups was 8%. The results are presented in Figure 8. The number near each point is the number of students in the group (the line that connects the dots does not have statistical meaning and is only there to guide the eye.)

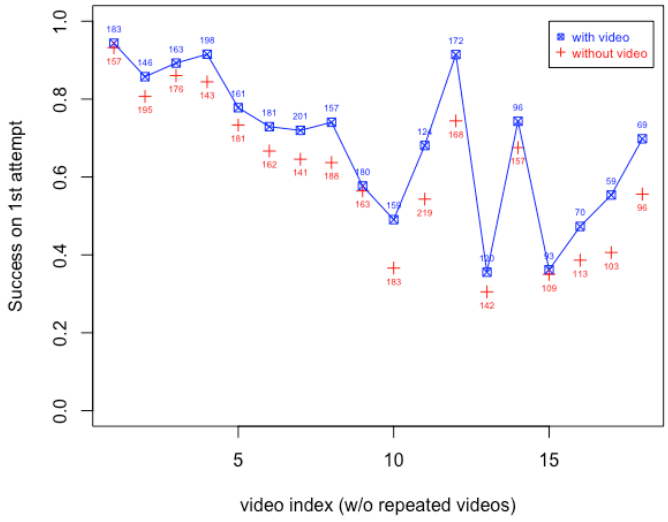

Figure 8: Between-group comparison

Regression Analysis

As described in Subsection 4.2, to evaluate the effect of watching the cartoon while adjusting for student and question parameters, we fitted the following logistic regression function:

$$Log\ Odds\ [CFA(Si,Qj)] = B0 + B1 * Si\_ability + B2 * Qj\_difficulty + B3 * Si\_watched\_Vj$$

The model was fitted using the standard *glm* function in *R*. Table 1 presents the resulting estimate (*B*), the standard error (*SE*), *Z* statistics, and the *p-value*, for the fitted model. A low *p-value* indicates that we can reject the null hypothesis that the corresponding term equals zero (no effect). Here, the *p-value* is very low for all the terms, which means that their effect is statistically significant. To determine whether this model provides a better fit than a reduced model without the watched covariate, we ran a Likelihood Ratio test (using *lrtest* from *R*'s *lmtest* library). We obtained a $p\text{-}value < 0.001$ and rejected the null hypothesis in favor of the model with the *watched* parameter.

Table 1: Results for the Logistic Regression (the resulting parameters for the watched video covariate)

term	B	SE	Z	*P*
(Intercept)	0.55	0.05	10.33	*<0.001****

watched	**0.35**	**0.07**	**4.82**	***<0.001******
ability	1.21	0.04	27.05	*<0.001****
difficulty	0.77	0.04	18.80	*<0.001****

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Let us now focus on the estimate (B) of watching. $B>0$ indicates a positive relationship between the covariate and the Log Odds. As shown in Table 1, after controlling for the student and question effect, watching the cartoon is associated with a 0.35 increase in the Log Odds ratio. Translating the Log Odds into Odds Ratio, e^B, provides a more interpretable measure of this relationship. For $B=0.35$, the Odds Ratio is 1.42. This means that for students who watch the cartoon, the Odds Ratio increases by 42%.

Turning the Odds Ratio into probability means that a student who has a 50% chance of solving the next question correctly *without* seeing the cartoon would have a 59% chance of getting the question correct if the cartoon was watched (an Odds Ratio of 1:1 translates into $p = 1/(1+1)=0.5$; 1.42:1 translates into a probability of $1.42/(1+1.42)=0.59$).

Discussion

This paper is based on a pilot project that focuses on developing Intelligent Tutoring Systems (ITSs) that cover topics within the K12 curriculum and that can offer a personalized learning experience. As part of this endeavor, we conducted educational data mining research that aims both to improve the pedagogy of the *specific* ITS, and to derive *general* instructional design principles for future ITSs that will cover additional topics.

The specific study presented here aimed to identify the effectiveness of educational cartoons, which were used within an ITS that teaches fractions to 4th graders, as a means for introducing mathematical concepts. This research relies on the fine-grained clickstream data that the ITSs collect, on Meta-data that describe the relationships between the entities (questions, pages, and cartoons, among others), and pedagogic Meta-data that map the activities into a taxonomy that models the concepts and skills that the students should master.

The rationale for using animated cartoons, and not texts or videos of a human teacher, is to make the learning experience more engaging and fun as well as to increase students' motivation to watch the cartoons, and increase their attention while watching them. Thus, the first issue that we examined is the level of participation, namely, to what extent are students actually watching the cartoons. The results (Figure 2) show that, on average, 75% of the students opened each video, and that about half of them watched each cartoon entirely. Since students were not obligated to see the cartoons, we interpreted these numbers as considerably high. The trend is also stable over time. The large dropout was evident in the repeated cartoons (Cartoons 5 and 9 in Figure 2). This suggests that students paid attention (thus, they immediately noticed the repeat), and that the decision to watch the cartoon was intrinsic (if it was extrinsic, we would expect to see more or less the same level of watching, regardless of whether the students felt that the cartoons would be valuable for them).

A point of interest for further design is the optimal length for educational cartoons. Guo et al. (2014) reported 6 minutes as an optimal video length for MOOCs. We deal with much younger students, so it is reasonable to assume that we would see a shift towards a shorter length. Our findings show that the level of dropout was highest in the first 10 seconds, then decreased, and then started to rise after about 90 seconds. However, further research is required to determine the optimal video length. (We note that 90 seconds is quite close to results obtained in a very different context – commercial ads – by the video analytics company Wistia. Based on 1.3 billion play events, they found that 2 minutes is a 'tipping point' after which viewers' engagement tends to drop [7]).

Regarding individual preferences to watch/skip the videos, we did not find any strong evidence that this is a 'stable' trait (in the sense of a 'learning style'). The findings (Figure 3) show that the ~20% who do not open each of the cartoons are not the same students who avoided most of the videos.

Interestingly, we did observe appreciable gender differences, with larger dropout rates occurring among male students (Figures 6 and 7). On average, male and female students tended to open about the same number of cartoons. However, on average, males watched entirely 53% of the cartoons to which they had access, whereas for females this rate was 61% (a relative increase of 15%). This is in line with the anecdotal report that we received from the field instructor of the single-gender male/female classes, who reported that students in the male class said that they want to progress quickly; thus, they do not want to spend time watching the cartoons.

Next, in studying the view patterns, we examined the relationship between watching the cartoons and success with related questions. The first model that we used (Subsection 5.2; the results are presented in Figure 8) shows that per cartoon, the group of students who watched the cartoon tended to perform better (an absolute mean

[7] https://wistia.com/blog/optimal-video-length

gain of +8%) than the ones who did not watch it. Although the difference is not dramatic, it is significant (it translates to about 0.5 standard deviation).

The main threat to the validity of the between-groups comparison is the pre-treatment difference between the groups. The logistic regression model that we used (Subsection 5.2; the results are presented in Table 1) adjusts for the individual and the question level, parameterized using ability and difficulty (for students and questions, respectively). This model uses a local measure of learning – the first question after the cartoon. The results were qualitatively similar to the first model: The model predicts that students who did not watch the cartoon will have a 50% chance of getting the item correct and will have a 58% chance if they watched the cartoon (an absolute increase of 8%, as before).

The fact that two different analyses, which use a different measure of learning as the outcome, yield a similar result, triangulates the results. In addition, the fact that the first model, which evaluates the performance for the entire sequence, also shows that the knowledge gap induced by not seeing the cartoon is not closed during the sequence by only solving questions (in fact, we also tried a model that evaluates the learning only for the last question, which yielded similar results).

Overall, we observed a positive relationship between watching the cartoons and the likelihood to succeed on related items. Because the cartoons introduce the concepts and explain them, followed by activities that expand on the topic, we believe that this relationship is causal.

Of course, threats to the internal validity of the results exist, indicating a causal effect. As is always the case with quasi-experimental research design, the results could be affected by confounding covariates that were not measured. For example, the logistic regression model adjusts for differences between students, but it does not adjust for temporal differences within students (e.g., ups and downs in students' motivation, which might affect both the decision to watch the video and the performance on the subsequent questions).

Conclusions and Future Research

This study presents the results of educational data mining research that aims to evaluate the use and effectiveness of animated cartoons as a means to explain mathematical concepts and problem-solving strategies in an Intelligent Tutoring System for 4th grade Math students.

We found that students tended to watch the cartoons even though they could proceed without watching them and that dropout was relatively high in the first few seconds, then dropped, and then started to rise again after about 90 seconds of watching, and that male students tended to dropout earlier. We also found a positive correlation between watching a cartoon and success on related items, which we believe is causal.

This study is part of a larger effort aimed at developing data mining methodologies that i) provide instructional designers with learning analytics that can guide iterative improvement of existing content, and ii) provide insights into students' learning, from which more general design principles can be derived. To the best of our knowledge, this paper is the first to report such an analysis of high-resolution clickstream data in order to study the use of videos in K6 Intelligent Tutoring Systems.

In the future, we intend to extend our analysis and study the intersection between the content of the cartoons and help-seeking patterns. For example, it would be interesting to determine whether students return to specific areas within the cartoons when seeking help, and whether the consistent structure of the cartoons (opening->problem->explanation->closing) facilitates effective searching. It would also be interesting to dissect the behavior by student cohorts, and, for example, to study whether different performance groups gain different values from watching cartoons.

Acknowledgments

We would like to thank The Center for Educational Technology for giving us access to the data, and Hagar Rubinek, Yael Weisblatt, and Oren Harel for their help in conducting this research.

References

Alexandron, G., Zhou, Q., & Pritchard, D. E. (2015). Discovering the Resources that Assist Students in Answering Questions Correctly – A Machine Learning Approach. *Proceedings of the 8th International Conference on Educational Data Mining*, 520–523.

Barak, M., Ashkar, T., & Dori, Y. J. (2011). Computers & Education Learning science via animated movies : Its effect on students ' thinking and motivation. *Computers & Education*, *56*(3), 839–846.

Bergner, Y., Colvin, K., & Pritchard, D. E. (2015). Estimation of ability from homework items when there are missing and/or multiple attempts. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*,

118–125.
Chen, Z., Chudzicki, C., Palumbo, D., Alexandron, G., Choi, Y.-J., Zhou, Q., & Pritchard, D. E. (2016). Researching for better instructional methods using AB experiments in MOOCs: results and challenges. *Research and Practice in Technology Enhanced Learning*, *11*(1), 9.
Dalacosta, K., Kamariotaki-paparrigopoulou, M., Palyvos, J. A., & Spyrellis, N. (2009). Computers & Education Multimedia application with animated cartoons for teaching science in elementary education. *Computers & Education*, *52*(4), 741–748.
Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement. *Proceedings of the First ACM Conference on Learning @ Scale Conference - L@S '14*, 41–50.
Kim, J., Guo, P. J., Cai, C. J., Li, S.-W. (Daniel), Gajos, K. Z., & Miller, R. C. (2014). Data-driven interaction techniques for improving navigation of educational videos. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology - UIST '14*, 563–572.
Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, Z., & Miller, R. C. (2014). Understanding in-video dropouts and interaction peaks in online lecture videos. *Proceedings of the first ACM conference on Learning@scale*., 31-40
Krishnan, S. S., & Sitaraman, R. K. (2013). *Understanding the effectiveness of video ads*. In *Proceedings of the 2013 conference on Internet measurement conference - IMC '13*, 149-162.
Machardy, Z., & Pardos, Z. A. (2015). Evaluating The Relevance of Educational Videos using BKT and Big Data. *Proceedings of the 8th International Conference on Educational Data Mining*, 424–427.
Oktay, H., Taylor, B. J., & Jensen, D. D. (n.d.). (2010). Causal Discovery in Social Media Using Quasi-Experimental Designs. *Proceedings of the First Workshop on Social Media Analytics,* 1-9.
Roll, I., Baker, R. S. J. d., Aleven, V., &Koedinger, K. R. (2014). On the Benefits of Seeking (and Avoiding) Help in Online Problem-Solving Environments. *Journal of the Learning Sciences*, 23(4), 537–560.