CRESST REPORT 812

USE OF A SURVIVAL ANALYSIS TECHNIQUE IN UNDERSTANDING GAME PERFORMANCE IN INSTRUCTIONAL GAMES

FEBRUARY, 2012

Jinok Kim

Gregory K.W.K. Chung

National Center for Research on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

Use of a Survival Analysis Technique in Understanding Game Performance in Instructional Games

CRESST Report 812

Jinok Kim and Gregory K.W.K. Chung
CRESST/University of California, Los Angeles

February, 2012

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305C080015.

The findings and opinions expressed in this paper do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

To cite from this report, please use the following as your APA reference: Kim, J., & Chung, G.K.W.K. (2012). *Use of a survival analysis technique in understanding game performance in instructional games.* (CRESST Report 812). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

USE OF A SURVIVAL ANALYSIS TECHNIQUE IN UNDERSTANDING GAME PERFORMANCE IN INSTRUCTIONAL GAMES

Jinok Kim and Gregory K. W. K. Chung
CRESST/University of California, Los Angeles

Abstract

In this study we compared the effects of two math game designs on math and game performance, using discrete-time survival analysis (DTSA) to model players' risk of not advancing to the next level in the game. 137 students were randomly assigned to two game conditions. The game covered the concept of a unit and the addition of like-sized fractional pieces. The math content in the baseline version of the game focused on procedures and did not elaborate on the math topics. The experimental version of the game provided more conceptual instruction by emphasizing the underlying concepts in fractional addition. Students played the game for 30 minutes. DTSA was used to examine student game performance, and whether and how game performance relates to students' prior math knowledge and game experience. Students who played the experimental version of the game were less likely to fail in the game relative to students who played the baseline version of the game (odds ratio = 0.64). Students with higher prior knowledge of fractions also were less likely to fail in the game (odds ratio = 0.41), and students with more game experience were less likely to fail (odds ration = 0.58). The use of DTSA provided an analytical method to understand game performance and game process data. DTSA enabled examination of the game play progression of students with various characteristics over sequences of game levels.

Introduction

Instructional or educational games have been increasingly developed and implemented in various fields. Instructional games and simulations are frequently used in military training, medical training, and are becoming more popular in school settings as well (e.g., FAS, 2006; Hays, 2005; Sitzmann, 2011; Tobias & Fletcher, 2011). For example, curriculum units based on instructional games, simulations, and technology replaced entire traditional science curriculum in some small districts. The increasing use of games in school settings raises interest in what instructional games can and cannot do in combination with, or sometimes in place of, traditional education. Thus, research interests often lie in whether game-based learning is associated with increased engagement levels of students—and ultimately with positive learning gains.

Although there has been some literature on the relative effectiveness of the use of games versus the non-use of games, there has been scant research on the ways that games can support and facilitate education in school settings. For example, game-based learning in school settings might provide important mediums to instruction, such as serving as diagnostic assessment or even formative assessment (Black, Harrison, Lee, Marshall, & William, 2003; Black & William, 1998) in the context of traditional education.

Determining whether game-based learning in schools can serve as diagnostic or formative assessment might be tantamount to judging whether performance in game-based learning can represent students' underlying levels of knowledge in a domain targeted by an instructional game. Or more ambitiously, it is tantamount to judging whether game performance from game process data can represent students' learning progression (Briggs, Alonzo, Schwab, & Wilson, 2006; Wilson, 2009) in a targeted domain.

This report attempts to explore such possibilities by empirically investigating game performance, and whether, or to what extent, game performance can inform student understanding in the targeted domain. This report draws on data from an intervention study in which over 100 students played an instructional math game, *Save Patch,* which was designed to enhance students' knowledge of the concepts of unit, numerator, denominator, and fractional addition. The study was originally designed to see the relative effectiveness of different versions of the game. This report examines the original question of the study; in addition this report utilizes the data and illustrates a technique that can help shed light into the learning progression of students in the domain targeted in the game.

Specifically, we build on the general property of instructional games, which often consist of a sequence of levels, with higher levels tending to require additional knowledge or skills in domains of interest. Then we use the discrete-time survival analysis (DTSA) technique to understand student game performance, and whether and how game performance relates to various factors of interest. Survival analysis (Efron, 1988; Hosmer & Lemeshow, 1999) is a statistical method that is uniquely suited to deal with time-to-event data due to its unique capacity to make accurate inferences concerning event occurrences and their timing. Since it is a rare application of the survival analysis technique, despite the fact that survival analysis is a broadly utilized method in various disciplines, we need an analogue between game data and time-to-event data. In the framework of DTSA, performance in an instructional game is assessed by whether players "beat the game," or successfully pass through all levels of the game until the final level of the game; and also by when (or at which level) players failed in the game, if they did not beat the game. Thus, the event is defined by failing to advance to the next level of the game; the time is defined by the levels in the game.

This report focuses on and illustrates two ways in which DTSA is extremely useful in analyzing student game performance. First, DTSA may provide systematic ways to understand the learning progression of players with various backgrounds, by describing how learners progress to the more advanced levels in the game (for example, life tables and survival probabilities as will be illustrated below). The learning progression that the data show—both descriptively and via estimation—may be compared to the hypothesized patterns of learning progression (e.g., across subgroups, across different levels of background knowledge, and across different game designs).

Second, by using survival analysis techniques we can borrow the capability of handling censored observations, which allows us to draw sound inferences concerning relationships of game performance to various factors of interest. Censored observations are a general problem in studies that examine event occurrences and their timing, since it is often the case that studies include many subjects who did not experience the event and thus the timing is unobserved (i.e., censored). Instructional games usually aim at knowledge and skills in certain domains; thus, it is often the case that some learners will never experience failure to advance to the next level (which is the event in survival analysis of game performance). This is mainly because true instructional games typically deal with specific content areas and include bounded levels of knowledge or skills. It is rare that a single game is designed to reflect an entire domain of interest (e.g., algebra). Thus, for learners who reach the final level of a game, the underlying knowledge or skill levels of the learners can go beyond the levels comprising the game. In such cases, the true level of game performance can be thought of as unobserved or "censored."

Failure to take into account censored observations may yield biases. In the presence of a large number of censored observations, even simple statistical summaries such as means and standard deviations can be biased in the analysis of event occurrences and their timing (Singer & Willet, 2003). Thus, by borrowing the capability to take into account censored observations, the use of the survival analysis technique can help draw sound inferences concerning game performance. For example, in efficacy trials comparing different game conditions, in which there were appreciably differential proportions of censored observations between conditions, unlike other naive analysis, survival analysis can yield an unbiased estimate of the impact or comparison parameter.

In what follows, we first present a framework in which we discuss some of the background of survival analysis, and ways to apply survival analysis to data on game performance. Next, we illustrate this approach, drawing on data from a study of an instructional game, *Save Patch*, developed and designed to enhance understanding of mathematical knowledge in the domain of rational numbers/fractions. Then, we use DTSA techniques to answer various research questions, such as validation of learning progression, student characteristics associated with game performance, and relative effectiveness of different game designs on game performance, and then present the results. Lastly, we discuss implications of our findings in the context of using games in math learning and assessment.

Use of DTSA Techniques for Understanding Game Performance

In this section, we briefly introduce the survival analysis technique, and discuss the building blocks of the survival analysis technique in the context of understanding game performance so that the technique can be readily applicable to game performance data. We note that the types of games which we focus on in this paper are instructional games that consist of a sequence of levels, with higher levels tending to require more skills or knowledge in the domains of interest.

The survival analysis technique is a strand of statistical methods and has been broadly used in various fields. The method may be most well known in medical fields, for example, examining whether patients with certain diseases die or survive and, if they die of the disease, when they do. In sociology, the same technique is known as event history analysis and is applied to study life events such as if marriage or divorce occur, and if they do, when they occur. In economics where the technique is known as duration analysis, events such as unemployment and its duration (in other words, when the unemployed get hired since the time of unemployment) are examined.

The unique capacity of survival analysis is related to censored observations. Observations are referred to as “censored” when we do not observe the time when the event of interest occurs. Censoring arises mainly because some studies do not last long enough to observe the events of interest for some observations, because some observations are lost in follow-ups, or because the event never occurs for some observations. In addition, survival analysis allows us to readily accommodate estimating the relationships of time-varying predictors to event occurrences and their timing. Event occurrences can be related to time-varying predictors as well as other predictors whose values are constant over periods of interest. For example, when studying criminal recidivism, predictors that are constant across time—such as gender, age, and previous education level—may be related to the re-arrest of released inmates. At the same time, time-varying predictors (such as income status, employment status, or marital status) may also be a significant predictor of re-arrest, which may change over the period of the study.

Since survival analysis techniques specially deal with time-to-event data, a time metric that is a meaningful scale for the event of interest should be specified for analysis. Depending on the length of the unit in the time metric, the metric can be viewed as either discrete or continuous. Such division is not trivial in survival analysis since it divides which technique to use: continuous-time methods or discrete-time methods. DTSA is used, and is more appropriate, in settings where events can occur at regular, discrete points in time. In other settings where the division of either continuous or discrete is unclear, whether many observations have the same values in the timing of event occurrence may be a practical indication of viewing the time as a discrete metric (Allison, 1995; Singer & Willet, 2003). Continuous-time methods can encounter difficulties in estimation when many observations have the same exact timing of event occurrences.

Whether discrete or continuous, a primary interest in survival analysis techniques lies in the latent variable *T* that represents the uncensored time of event occurrence, which may or may not be observed. The estimation fully uses information from both observed cases and unobserved (censored) cases (see for example, Efron, 1988).

To apply the DTSA technique to understand performance in gameplay, we should develop some analogies between components in survival analysis and those in game performance. In instructional games with multiple levels that require increasing levels of skills or knowledge to pass the levels, we may think of game performance as two questions: (a) whether players finish the final level of the game, or "beat the game"; and (b) if they fail to beat the game, when (or at which level in the game) they do. This way of seeing game performance provides a direct link to survival analysis, which is often referred to as a whether-and-when test. Based on such a notion, we define the building blocks of survival analysis, that is, the event of interest, the time metric, and the beginning point in the time metric, in the following ways:

1. the event of interest is whether learners fail to advance to the next level;
2. the time metric is the levels in the game; and
3. the beginning of the time metric is the beginning level in the game.

One may think that, among this set of analogies, the most disparate component probably is the time metric. Although survival analysis has been broadly applied to study different types of events in various fields as noted earlier, the time metric has been real time units (e.g., minutes, hours, months, or years). To make survival analysis technique applicable to and meaningful in understanding game performance, there needs to be a conceptual reorientation to this technique so that the game levels can be time clocking variables.

When we define the time metric as the levels in the game, the latent variable *T*, which represents uncensored time of event occurrence in survival analysis, becomes the *uncensored levels where learners fail to advance to the next level*. What we want to measure when we say "game performance" may not be the actual last game level that learners played in the game, but rather underlying levels of knowledge or skills in the domain of the specific game. In this line, survival analysis provides a better estimate of game performance because of its capability of dealing with censored (unobserved) observations. When we think of game levels as the time metric, the choice of method in survival analysis should be discrete-time methods instead of continuous-time methods. Game levels are clearly distinct categories that depend on specific games, rather than units on continuums.

One major reason why the timing of event occurrence is censored in instructional game settings is that the true levels of knowledge and skills of some players can be out of the range of the game. For example, when asked to play games on a specific domain such as fractions, some of the students may have mastered the domain and finished the final level of the game, that is, "beat the game." For these students, the final level may be a good representation of their level of knowledge or skills in the domain—their true level may be the final level in the game. However, for other students, their true level of performance could be way beyond the final level in the game in the underlying continuum. These students can be seen as "censored" observations. More generally, censoring arises mainly because some studies do not last long enough to observe the events of interest for some observations, because some observations are lost in follow-ups, or because the event never occurs for some observations. The above students in game performance settings can be the observations that are censored because they will never experience the event of interest. In the presence of a substantial proportion of possible censoring, survival analysis is not only applicable but an essential technique to accurately understand game performance.

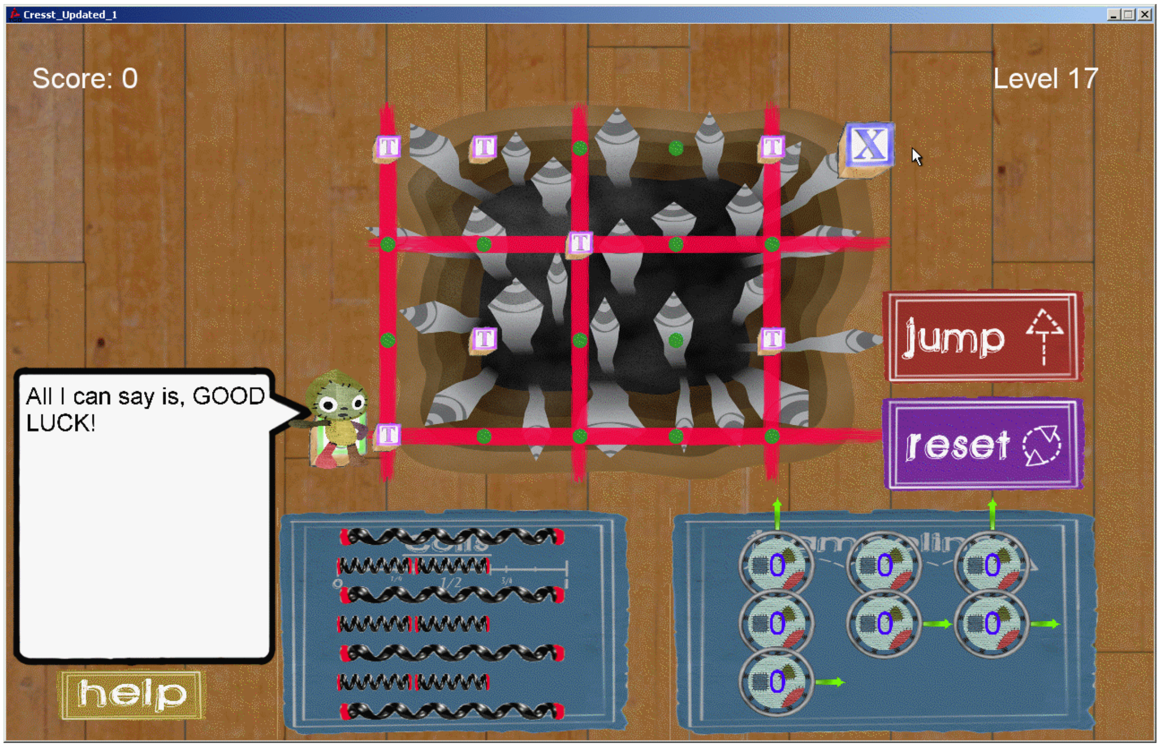
Methodology

In this study, we employ the DTSA technique to understand performance of over 100 students who played a prototype game, *Save Patch*. *Save Patch* is designed to enhance students' knowledge of the concepts of unit, numerator, denominator, and fractional addition. The game consists of 19 levels that increase in difficulty and knowledge demands.

Game Context

In collaboration with CRESST, USC's Game Innovation Lab designed and developed a prototype game that was used as a research testbed (see Figure 1). The game requires students to apply concepts underlying rational number addition. Students are presented with the challenge of helping the game character Patch jump over hazards to safely reach a destination. To do so, students place trampolines at various locations along a one- or two-dimensional grid. Each trampoline is made bouncy by the student dragging one or more coils onto the trampoline. A coil has a value that represents a whole unit or fractions thereof. The distance Patch will bounce is the sum of all coils (values) added to the trampoline. For example, if the student drags a coil of one unit onto a trampoline, that trampoline will cause Patch to bounce exactly one unit.

In *Save Patch*, one whole unit is always the distance between two lines (see Figure 1). This unit becomes the referent for coils of fractional bounce later on. Coils can be added to a trampoline to increase the distance Patch will bounce; however, only coils of the same fractional size can be added together. While any size coil can be placed on the trampoline initially, subsequent coils can only be added to the trampoline if they are the same size (i.e., have the same denominator).



All I can say is, GOOD LUCK!

Figure 1. Screen shot of *Save Patch*. One whole unit is always the distance between two red lines.

The game exploits an important property of real numbers—numbers can be broken into smaller, identical parts to facilitate addition, a process similar in both integer and fractional addition. An important game design goal was to make explicit the connections between integer addition and fractional addition. Moreover, the gameplay requires that players focus on the size of a unit when they are adding coils to a trampoline.

As gameplay proceeds, trampolines must be placed at distances along the grid that are fractional parts of the whole unit. In early game levels students are given the fractional unit coils. In later levels, students are shown how to break coils into fractional units. Because only coils that have identical units can be added together, students must be attentive to what the rational number means, to what units are being added, to what units are already on the trampoline, and to how they will break coils into different size pieces. So while students could add a coil that is 1/2 a unit to another coil that is 1/2 a unit, they are not allowed to add a coil that is 1/2 a unit to a coil that is a whole unit until the whole unit is broken into two 1/2-unit coils (i.e., 2/2). At the time all three of these coils are added to the trampoline, the trampoline would show that it had 3/2 (rather than 1 1/2) units of bounce. This notation is intended to reinforce both the meaning of addition and to reinforce the player's understanding of the meaning of rational numbers.

In *Save Patch*, the procedure for converting fractions of different sizes (i.e., fractions with different denominators) is not accomplished through multiplication. Rather, students are shown how they could break the coil into more pieces (each smaller in size) or fewer pieces (each larger in size), respectively. For example, a coil that was one whole unit could be broken into two halves, three thirds, and so forth. If the student used the same procedure with a 1/2 coil, then the coil broke into two fourths, three sixths, and so on.

The Appendix shows an excerpt from the development of all levels in the game. (For a full version and for more details, see Center for Advanced Technology in Schools (2010.) As shown in the Appendix, as the levels advance in the gameplay, higher levels of the game tend to require additional knowledge or skills in mathematics and fractions than the previous levels. There are certain levels that require more knowledge of fractions. In Levels 6 and 7, the game requires further knowledge of concepts of fractions and the addition of fractions (e.g., fractions can be further broken into parts, knowledge specification 2.1.3), while in the previous levels (Levels 1 through 5) the game is more focused on the representation of one whole unit (although Level 5 shows a whole unit coil broken into two pieces). The next couple of levels require a similar type of math knowledge and skills as in Levels 6 and 7.

Starting at Levels 10 and 11, the gameplay requires knowledge of converting fractions and finding the least common denominator. Although some previous levels (from Levels 6 to 9) may also involve some knowledge of converting fractions, students can also solve problems by counting broken pieces. However, to pass Levels 10 and 11 successfully, students need more accurate knowledge of how to convert fractions and need to have mastered the associated skills to some extent.

Also, as can be seen in the Appendix, from Level 15 and on, in addition to the math knowledge and skills required in the previous levels, the game also asks students to figure out the path to solve the levels successfully. Cognitive load in these levels may significantly increase because players have additional cognitive tasks while exercising high levels of knowledge on fractions.

Game versions. Two versions of the game were developed for testing. The initial version of the game was developed by the game developers and then modified by a subject-matter expert to emphasize the mathematics concepts. One version, considered the "baseline" version, contained the minimal amount of instruction needed to play the game. The math instruction focused on procedures and did not elaborate on the math topics. In addition, only minimal help was provided when errors were committed, and the game did not remind students about the math they used at the end of each level. In contrast, the experimental version provided more conceptual instruction by emphasizing concepts such as the whole units in the level (the distance between the red lines), that only pieces of the same size can be added together, and the importance of the unit and piece size.

Research questions. Drawing on the data from a study in which a total of 137 ninth grade students played the game *Save Patch* for about 30 minutes, this paper is primarily concerned with game performance, individual differences in game performance, and relative effects of different game designs on game performance. Our specific questions are as follows:

1. Which DTSA model best describes the event-to-time data in our sample? How well does the model fit to the data, and also agree with sequences of underlying math knowledge in the game?
2. Which students are at greater risk of not being able to advance to the next levels in the game? Why do some students fail to advance to the next levels in the game?
3. Do different game versions affect student performance? More specifically, is there a significant effect of the experimental version of the game relative to the baseline version of the game on student game performance (i.e., the risk of failing in the game), controlling for important student characteristics?

Sample and Variables

Our study measured a number of student characteristics, ranging from their demographics and pretest scores in the domain of fractions to attitudinal and motivation measures and numerous behavioral measures concerning use of games and computers (see Baker, Chung, and Stigler, 2010, for the survey instrument and actual items as well as their descriptive statistics). An exploratory analysis indicated several student characteristics as potential correlates of game performance, such as the pretest scores, gender, self-reported amount of time spent on video game play per week, self-reported levels of game ability, and competitive learning preference. Table 1 presents descriptive statistics of only those variables that are potentially related to our outcome of interest, that is, game performance.

The two right panels in Table 1 present the descriptive statistics by game conditions. The data were collected through three administrations, and in each administration we randomly assigned individual students to either the experimental version or baseline version of the game. As can be seen, the means and standard deviations are extremely close between groups playing the two game versions. For example, of the 134 participants who reported gender, 58% were female in the baseline condition, and 56% were female in the experimental condition. An exploratory *t* test showed that there is no significant difference in any of the pretreatment characteristics shown in Table 1 between students assigned to the two conditions. Given the random assignments and the above results, inferences concerning the relative effect of the experimental version of the game to the baseline version of the game presumably have a strong internal validity.

Table 1

Descriptive Statistics for the Entire Sample and by Game Version

	Sample			Baseline			Experimental		
Variable	*N*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Pretest score on core items[a]	137	6.34	3.39	68	6.31	3.36	69	6.36	3.44
Frequency playing video games[b]	136	2.42	1.05	68	2.32	0.98	68	2.51	1.11
Length of time playing video games[c]	136	4.32	1.20	68	4.24	1.22	68	4.40	1.17
Level of ability with video games[d]	136	3.36	0.98	68	3.26	1.02	68	3.46	0.94
Frequency of computer use (in hours a week)	130	8.77	13.16	65	9.16	15.14	65	8.38	10.94
Attitude scales									
Self-belief: control expectation[e]	136	3.02	0.66	68	3.02	0.68	68	3.02	0.64
Self-belief: perceived self-efficacy[e]	136	2.73	0.69	68	2.76	0.70	68	2.69	0.68
Learning preferences: competitive[f]	136	2.84	0.74	68	2.80	0.73	68	2.87	0.76
Learning preferences: cooperative[f]	136	2.95	0.69	68	2.83	0.69	68	3.07	0.66
Motivation: interest in math[f]	136	2.29	0.84	68	2.20	0.80	68	2.37	0.87
Self-belief: math[f]	136	2.48	1.02	68	2.36	0.96	68	2.61	1.06

[a] Maximum score was 11.
[b] 1 = none, 2 = 1 to 2 hours a week, 3 = 3 to 6 hours a week, 4 = more than 6 hours a week.
[c] 1 = never, 2 = less than 1 year, 3 = 1 to 2 years, 4 = 2 to 3 years, 5 = more than 3 years.
[d] 1 = poor, 2 = fair, 3 = average, 4 = above average, 5 = far above average.
[e] 1 = almost never, 2 = sometimes, 3 = often, 4 = almost always.
[f] 1 = disagree, 2 = disagree somewhat, 3 = agree somewhat, 4 = agree.

Statistical Methods and Models

This section describes specific analysis and fitted models for each research question of this study. We define “event” as *failing in a certain level* or *failing to advance to the next level during the time allowed for the gameplay* (30 minutes after students have the opportunity to learn how to play the game; i.e., 30 minutes after passing Level 1). Our timing is clocked as levels of the game advance. In other words, in our study, “events occur” in settings where students fail to advance to the next level of the game in the given gameplay, and “when” they occur is measured by the levels of the game at which students fail to pass in the gameplay.

Describing learning progression and game performance. A useful tool to describe event occurrence over time is a life table. The table will first show levels that are available in the game ranging from Levels 1 to 19. For each level *i* from 0 to 19, it provides three types of numbers that are important in describing event-to-time data: (a) number of students who played (who are at risk) at the beginning of level *i* (n_i); (b) number of students who failed during level *i* (s_i); and (c) number of students lost to follow-up during level *i* (s'_i). The group of students who played at the beginning of each level is referred to as a risk set in survival analysis, and is equivalent to the number who played at a previous level subtracted by the number who failed or lost at the previous level: $n_{i+1} = n_i - s_i - s'_i$.

A life table also provides two kinds of probabilities, first, the probability of students who failed during level *i* among the students who played in the beginning of level *i*, and second, the probability of students who are still playing at the end of level *i*. The first type of probability is referred to as a hazard probability or hazard rate, which is in fact a conditional probability. The hazard probability, h_i, is:

$h_i = Pr$\{student fails during the *i*th level | student passes until the beginning of *i*th level\},
which can be obtained in the life table by $h_i = s_i / n_i$. (1)

The second type of probability is referred to as a survival probability. Formally, the definition is the probability that a student does not fail during the first *i-1* levels and thus plays at least until the beginning of the *i*th level. Formally, the survival probability G_i is:

$G_i = \prod_{0 \le j < i}(1 - h_j)$. Thus, by definition, $G_0=1$.

Fitting DTHMs to describe learning progression and game performance patterns. To obtain a base model that summarizes time-to-event patterns from our data, we fitted various discrete-time hazard models (DTHMs). The most general model in DTHMs is to pose the logit or log odds of hazard, which is defined in Equation 1, as a function of dummy variables indicating each time period in the study. This model is completely general and should yield the best fitting model in representing the effect of time, since the fitted values actually replicate hazards that are in the sample (as shown in life tables such as Table 2). Note that the model is not estimable when there is a time point where no event occurs (Allison, 1995). Our data turned out to be this case since we have a number of time points where no students failed. Also, the downside of this completely general model is noted. They can lack parsimony in settings where there are many time periods and can also reflect erratic patterns that may be nothing but sampling variability (Efron, 1988; Fahrmeir & Wagenpfeil, 1996; Singer & Willet, 2003).

Therefore, in this study, we pursue models that are parsimonious but still closely describe and summarize important patterns from the time-to-event data. In the DTSMs, (logit) hazard is predicted by a *Time* predictor as if they were a continuous variable, and higher orders (polynomials) of the *Time* predictor represent the time trends and potential curves in time trends. However, these are still in a discrete time framework so we constrain the fitted values and the interpretations in the range of *Time*'s observed values, which is from 1 through 19 (see for example, Efron, 1988; Singer & Willet, 2003, pp. 408–419).

Specifically, the following DTSMs are fitted. The models of the logit hazard probability h_i are a function of *Time* or additionally increasing orders of *Time*. The last model is completely general.

$$Logit\ h_i = \alpha_0 \tag{2a}$$

$$Logit\ h_i = \alpha_0 + \alpha_1\ Time_i \tag{2b}$$

$$Logit\ h_i = \alpha_0 + \alpha_1\ Time_i + \alpha_2\ Time_i^{\ 2} \tag{2c}$$

$$Logit\ h_i = \alpha_0 + \alpha_1\ Time_i + \alpha_2\ Time_i^2 + \alpha_3\ Time_i^{\ 3} \tag{2d}$$

and where *Time* represents game levels that range from 0 to 19, but is centered around 7; and I_j is a binary indicator of level j (j = 1..19).

The models assume a different representation of the effect of *Time*. The *Constant* model in Equation 2a poses that the hazard of failing in a level is not related to which level a student is. The *Linear* model in Equation 2b poses that the logit hazard increases or decreases steadily as the level increases. The following models each include polynomials of *Time*, for example, the *Quadratic* model includes a second order polynomial (Equation 2c), and the *Cubic* model includes a third order polynomial (Equation 2d). We also try to fit models with up to fourth and fifth order polynomials, but for these models the maximum likelihood (ML) estimation did not yield convergence.

In posing DTSMs that closely fit to the data, the original design of the game *Save Patch* raises an important issue. To summarize from the description of all levels in *Save Patch* (see the Appendix and discussions around it), Levels 6 and 7, 10 and 11, and 15 and 16 are where the game is designed to introduce new math knowledge or skills that are important in the domain. Therefore we should probably pose models that test the DTSMs that accommodate the introduction of the big ideas. To do this, we first examined in the life table the specific levels where students show sharp increases of hazard probability, or equivalently, significant sharp drops in the survival probability; and whether these coincided very well with the original game design. Next, we tested further DTSMs to see whether the sharp increases that emerge in an expected way in the gameplay are statistically significant and provide significantly better fit to the data.

If the peaks in the data are not sampling fluctuations, but rather abrupt changes in the time-to-event data pattern due to game designs, they should be modeled so that the study can further address those changes. In line with this, we fitted another DTHM which aims to incorporate the sharp or abrupt increases in hazard (or drops in survival probability) that smoothing functions cannot achieve alone. In the cubic function that was fitted above, we additionally included indicators of levels for 10 and 11, 15, and 16 to capture the abrupt changes. Specifically, we fitted a discrete-time hazard model with the following specification:

$$Logit\ h_i = \alpha_0 + \alpha_1\ Time_i + \alpha_2\ Time_i^{\ 2} + \alpha_3\ Time_i^{\ 3} + \alpha_4 I_{1011} + \alpha_5 I_{15} + \alpha_6 I_{16} \quad (3)$$

Student factors associated with game performance. In this section we examine the factors that are related to failing to advance to the next level in the game. In a DTSA framework, we probe possible correlates of the (logit) hazard by including and testing possible correlates in our base model (Equation 3). This helps investigate which students are at greater risk of failing in the game, and helps us understand why some students successfully pass through levels while others may struggle and fail in earlier levels in the game.

We have fit numerous models as part of exploratory analysis. Here we briefly outline the process used in exploratory analyses and summarize the results. Our study collected a large number of measures of student characteristics, including their pretest scores in the domain of fractions, various attitudinal and motivation measures, and various behavioral measures concerning use of games and computers (see Table 1). To locate a smaller set of variables to test in our discrete-time hazard models, we examined correlations between the last level in the game and various student characteristics. This exploratory analysis indicated several student characteristics as potential correlates of game performance, such as pretest scores, gender, self-reported amount of time spent on video game play per week, self-reported levels of game ability, and competitive learning preference.

All these characteristics identified as potential correlates of game performance were added into our base model (Equation 3), first one by one, and later in addition to other predictors that turned out to be significant. The specifications of some models we fitted with best set of predictors are presented below.

Model A:

$$Logit\ h_i = \alpha_0 + \alpha_1\ Time_i + \alpha_2\ Time_i^{\ 2} + \alpha_3\ Time_i^{\ 3} + \alpha_4 I_{1011} + \alpha_5 I_{15} + \alpha_6 I_{16}$$

Model B:

$$Logit\ h_i = \alpha_0 + \alpha_1 Time_i + \alpha_2\ Time_i^{\ 2} + \alpha_3\ Time_i^{\ 3} + \alpha_4 I_{1011} + \alpha_5 I_{15} + \alpha_6 I_{16} + \alpha_7\ Mathpretest_i$$

Model C:

$$Logit\ h_i = \alpha_0 + \alpha_1\ Time_i + \alpha_2\ Time_i^{\ 2} + \alpha_3\ Time_i^{\ 3} + \alpha_4 I_{1011} + \alpha_5 I_{15} + \alpha_6 I_{16} + \alpha_7\ Gameplay_i$$

Model D:

$$Logit\ h_i = \alpha_0 + \alpha_1 Time_i + \alpha_2\ Time_i^{\ 2} + \alpha_3\ Time_i^{\ 3} + \alpha_4 I_{1011} + \alpha_5 I_{15} + \alpha_6 I_{16} + \alpha_7\ Mathpretest_i + \alpha_8\ Gameplay_i \qquad (4)$$

In the above equations, $Pretest_i$ is the sum of items student i got right, which ranges from 0 to 11 (see Table 1 for the mean and standard deviation); $Gameplay_i$ is the weekly amount of video game play of student i, and ranges from 1 to 4, with *none* coded as 1, *1-2 hours* coded as 2, *3-6 hours* coded as 3, and *6 hours or more* coded as 4 (see Table 1 for descriptive statistics of the sample); and other variables are coded in the same way as the previous sections.

Relative effectiveness of different versions of the game on student game performance. This section examines whether there was a significant effect of playing the experimental version of the game relative to playing the baseline version of the game on student game performance (i.e., the risk of failing in the game), controlling for important student characteristics. To examine whether there is a significant difference in the game performance between the two versions, we fitted a DTHM with the following specification:

$$Logit\ h_i = \alpha_0 + \alpha_1 Time_i + \alpha_2\ Time_i^{\ 2} + \alpha_3\ Time_i^{\ 3} + \alpha_4 I_{1011} + \alpha_5 I_{15} + \alpha_6 I_{16} + \alpha_7\ Mathpretest_i + \alpha_8\ Gameplay_i\ + \alpha_9\ Gameversion_i \qquad (5)$$

where *Gameversion* is a binary indicator variable of the experimental version of the game (i.e., experimental version coded as 1 and control version coded as 0). Other variables are defined and coded in the same way as in the previous analyses. The key coefficient of this model is α_9, which represents the expected difference in the logit hazard probability between the experimental versus baseline versions of the game controlling for other variables in the model. Even in cases where the randomized sample shows good balances in various pretreatment characteristics we included the two student characteristics in the model as in Equation 4. Such inclusion presumably removes possible imbalances remaining in the sample and increases the statistical power of analysis.

Results

Describing Learning Progression and Game Performance

Table 2 presents the life table for this study. The first column shows levels that are available in the game ranging from Levels 1 to 19. A total of 137 students started to play the game and they were given about 30 minutes after they got used to the game (after they played the first level of the game). For each level *i* from 0 to 19, the second to fourth columns show three types of numbers that are important in describing event-to-time data: (a) number of students who played (who are at risk) at the beginning of level *i*; (b) number of students who failed during level *i*; and (c) number of students lost to follow-up during level *i*. In Table 2, it can be seen that all 137 students succeeded in passing through levels until Level 6. However, from Levels 7 to 9, a few students failed to pass the levels, one, two, and four students respectively. The number of students who fail to pass the levels sharply increases to two-digit numbers in Levels 9 and 10.

In Level 12 of the game, 93 students were left to play the game. Among these students, only a few to several students failed to pass Levels 12, 13, and 14. The number of students who failed to pass again sharply increases in Levels 15 and 16. In Level 17, only 34 students were left to play. Among these students, 15 students reached the last level of the game during the duration of the study, and among them 8 students actually passed the last level or finished the game. As the fourth column shows, for the 8 students, the event (failing during level *i*) never occurred and thus these students are censored.

The hazard probabilities and the survival probabilities also succinctly show the above described patterns. The hazard probabilities show a large increase, or equivalently, the survival probabilities show an abrupt decrease in the above mentioned levels.

Table 2

Life Table Describing the Levels in the Game (*N* = 137)

	Number			Proportion	
Level	Played at the level	Failed during the level	Censored at the end of the level	Failed to pass during the level	Still playing at the end of the level
0	137				1.0000
1	137	0	0	0.0000	1.0000
2	137	0	0	0.0000	1.0000
3	137	0	0	0.0000	1.0000
4	137	0	0	0.0000	1.0000
5	137	0	0	0.0000	1.0000
6	137	0	0	0.0000	1.0000
7	137	1	0	0.0073	0.9927
8	136	2	0	0.0147	0.9781
9	134	4	0	0.0299	0.9489
10	130	19	0	0.1462	0.8102
11	111	18	0	0.1622	0.6788
12	93	1	0	0.0108	0.6715
13	92	2	0	0.0217	0.6569
14	90	6	0	0.0667	0.6131
15	84	27	0	0.3214	0.4161
16	57	24	0	0.4211	0.2409
17	34	6	0	0.1765	0.1984
18	30	15	0	0.5000	0.0992
19	15	7	8	0.4667	0.0529

Research Question 1: Which DTSA model best describes the event-to-time data in our sample?

Table 3 presents the fit statistics of the models that test different functions of Time in Equations 2a to 2d. From the results shown in Table 3 we chose to use the Cubic model as the best parsimonious model. The model fit is increasingly better based on the AIC and deviance statistics, from the Constant, Linear, Quadratic to Cubic model. Figure 2 displays the fitted logit hazard probabilities based on the models with increasing polynomials (Constant, Linear, Quadratic, and Cubic), as well as the sample probabilities as a comparison. It is apparent that the Cubic function does the best job capturing the rapid increase in the sample logit hazard probability in the first several levels. Also, although the logit hazard probability increases at a decreasing rate after the middle levels, there seems to be a little shift of direction in this pattern in later levels: the probability increases at a faster rate again in Level 15 and on. Figure 2 also shows that the Cubic model is the only one that can reflect such shifts among all the models we fitted above.

Table 3

Goodness-of-Fit Statistics of Four Models Representing Time-to-Event Patterns

Models	AIC	Deviance	Difference
Constant	969.845	967.845	
Linear	720.689	716.689	251.156
Quadratic	719.282	713.282	3.407
Cubic	713.04	705.04	8.242

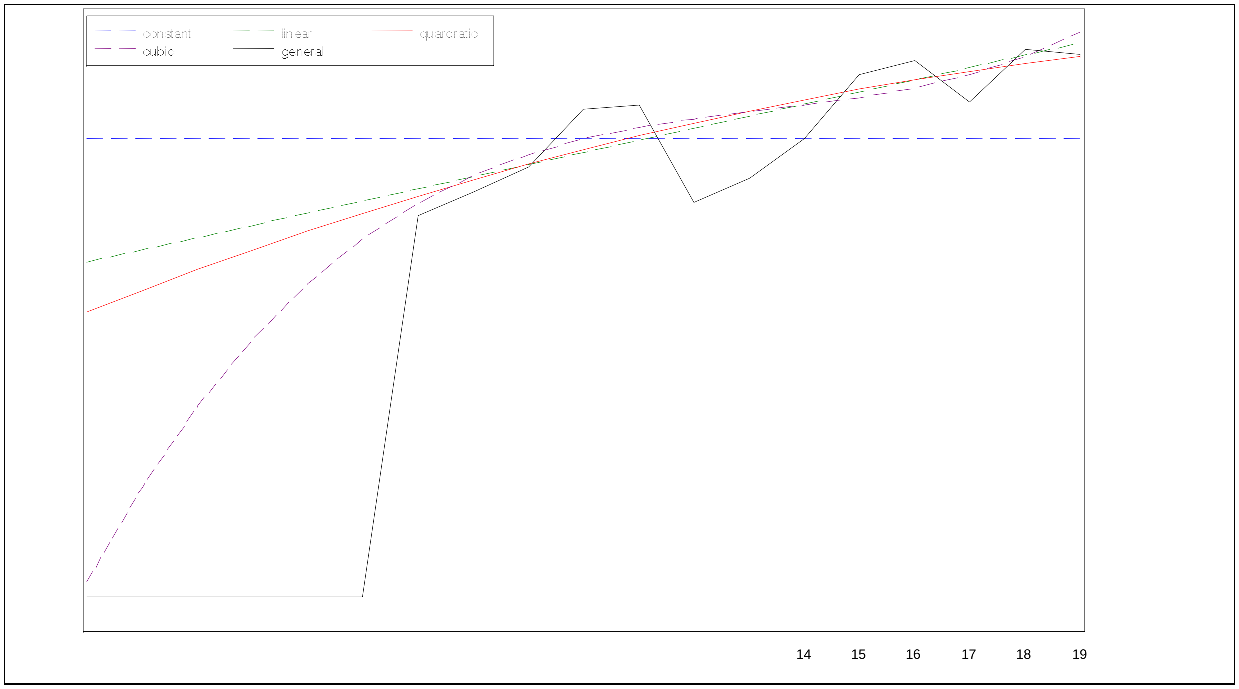


Figure 2. Fitted logit hazard probabilities based on the Constant, Linear, Quadratic, and Cubic models superimposed on the sample values.

Figure 3 displays the estimated hazard (the upper panel) and survival probabilities (the lower panel) against all levels in the game based on the Cubic model, superimposed on sample values. Although the cubic model smoothes the curve and replicates well the overall trend across all levels, there are some peaks in the data that may not be fitted from any smoothing functions. As shown in the life table (see Table 2) and also in the figure shown below, there are sharp increases in the hazard probabilities in Levels 10 and 11 and again in Levels 15 and 16. Thus, we fitted another DTHM which aims to incorporate the sharp or abrupt increases in hazard (or drops in survival probability) by including indicators of levels for 10 and 11, 15, and 16 to capture the abrupt changes, as shown in Equation 3.

Figure 4 displays the fitted hazard (the upper panel) and survival probabilities (the lower panel) based on the model that additionally included indicators of four levels (Equation 3), overlaid on those based on the Cubic model. One can see the close approximation of the new model to the sample probabilities. With visual comparison only, it seems apparent that the cubic function with level indicators in the above equation represents the time-to-event pattern in our data appreciably better, as compared to the cubic smoothing function alone (the Cubic model in Table 3). This conclusion is also evident from the results on goodness-of-fit statistics. The cubic function with level indicators yielded the deviance of 636.885 and AIC of 650.885, which is significantly better (*p* value < .0001) than the Cubic model in Table 3 (Deviance = 705.040; AIC = 713.040). Therefore, to answer Research Question 1, we choose the model with a cubic function and the level indicators to represent the trends over levels in this study. We refer to this model in Equation 3 as the “base model” in the rest of this report.

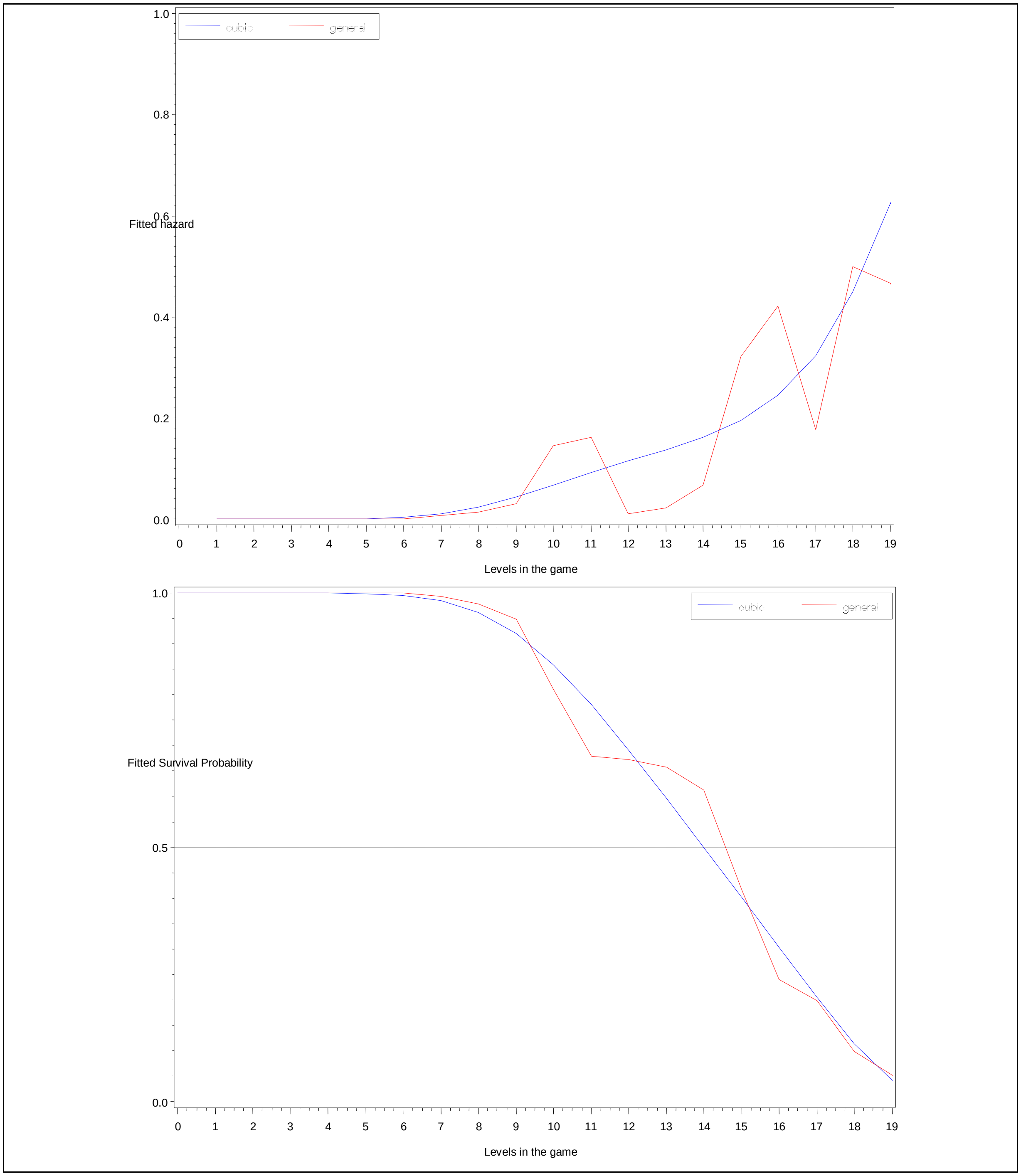


Figure 3. Fitted hazard probabilities (upper panel) and survival probabilities (lower panel) based on the Cubic model, superimposed on the sample hazard and survival probabilities.

Based on our base model (represented as a solid line in Figure 4), the risk of failing in the game increases rapidly in Levels 10 and 11 to 0.14 and 0.17 and decreases again in the next couple of levels. Then the risk increases in Level 15 and above to more than 0.20, and peaks at Levels 16 and 19 to higher than .40.

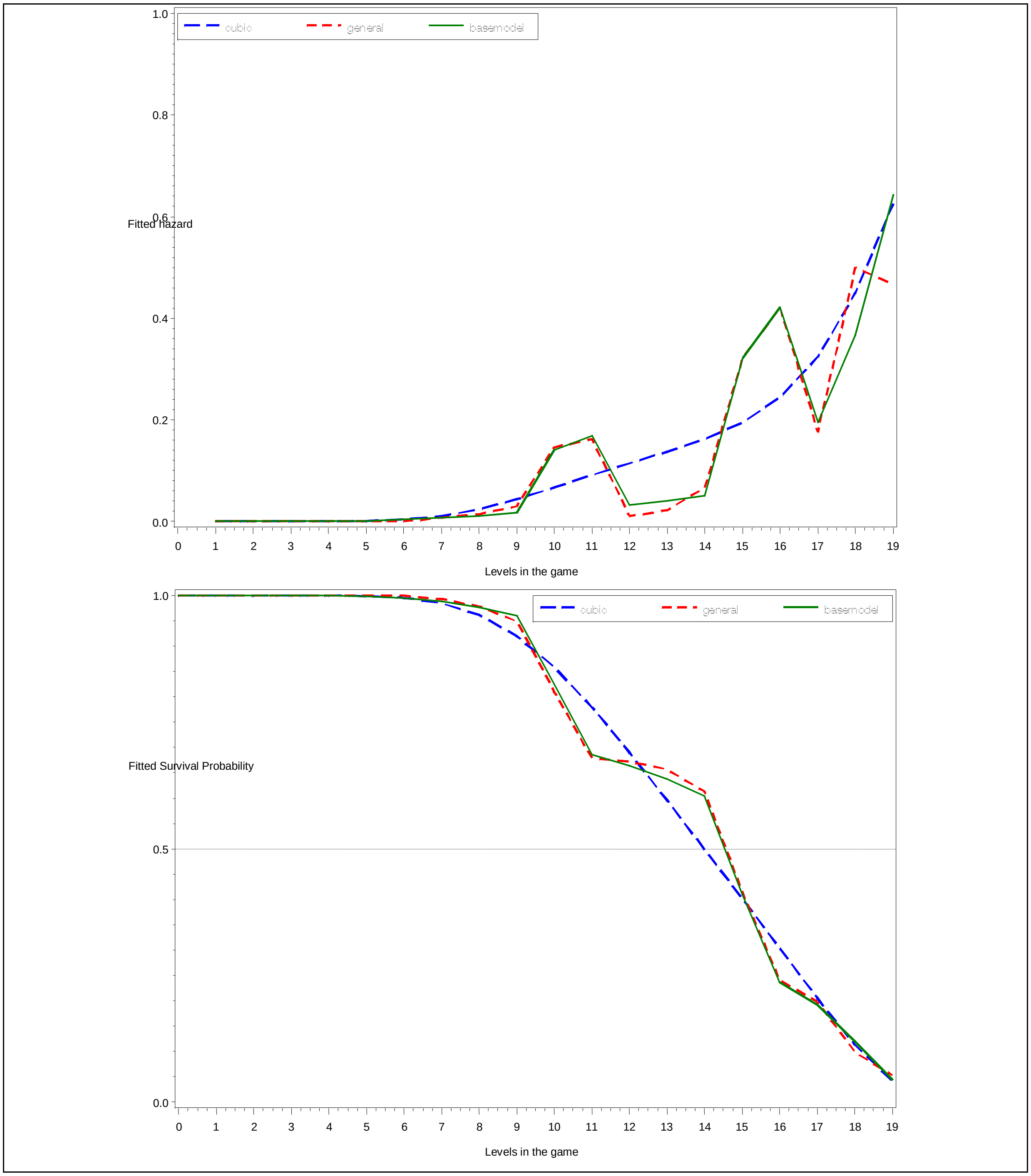


Figure 4. Fitted hazard probabilities (upper panel) and survival probabilities (lower panel) based on the Cubic model (Equation 2d) and the base model (Equation 3), superimposed on the sample hazard and survival probabilities.

Research Question 2: Which students are at greater risk of not being able to advance to the next levels of the game?

All the characteristics identified as potential correlates of game performance are added into our base model (Equation 3), first one by one, and later in addition to other predictors that turned out to be significant. The results overall indicate that, once two variables are in the model, which are student pretest score and the amount of time for video game play per week, no other variables emerged as significant predictors of the (logit) hazard of not advancing in the game.

For example, female students were more likely to fail in the game when the gender variable was added to the base model. Even when pretest scores were controlled for, gender still turned out as significant with the estimated odds of 1.60 (*p* value = 0.04). However, gender was not significant anymore (*p* value = 0.57) once the amount of time for video game play was included in the model.

Thus, the best set of predictors of hazard was found to be pretest scores in the domain of fractions and the amount of time students reported spending on video game play per week. Table 4 presents results of fitting various models focusing on the best set of predictors. Based on the goodness-of-fit results, Models B and C including math pretest scores and amount of time for video game play respectively indicated significantly better fits than the base model (*p* value < .0001). Model D including both variables indicated a significantly better fit than both Models B and C (*p* values < .0001). Now we turn to the estimates from Model D, which is our tentative final model for explaining individual differences in game performance. As shown in Table 4, the coefficient of *Pretest* is -0.90. Its negative sign indicates that students who came with higher prior levels of knowledge on fractions are less likely to fail in the game. When the coefficient is converted to an odds ratio metric, the odds ratio associated with the pretest score variable is 0.41. Thus, for students whose pretest scores are one standard deviation higher, the estimated odds of failing in the game tend to be 41% of the estimated odds of students whose scores are one standard deviation lower. Note that the pretest variable is standardized before being added to the model, which means that one unit is equivalent to one standard deviation (*SD*). For example, if the estimated odds is 0.12 (the hazard rate of 0.11) for a student whose pretest score is around the sample mean, for a student whose pretest score is 1 *SD* higher than the sample mean but whose other characteristics are comparable, the estimated odds is 0.05 (the hazard rate of 0.05), which is the product of 0.12 and 0.41.

Table 4

Results of Fitting Four Discrete Time Hazard Models to the Data (*n* = 136)

	Model A	Model B	Model C	Model D
Parameter estimates and standard errors				
Intercept	-12.5130	-12.4426	-10.5844	-10.8058
	(4.8192)	(4.6584)	(4.6709)	(4.4717)
Time	2.1062	1.8695	1.9161	1.6840
	(1.4069)	(1.3715)	(1.3726)	(1.3249)
Time^2	-0.1797	-0.1571	-0.1612	-0.1386
	(0.1272)	(0.1248)	(0.1249)	(0.1213)
Time^3	0.0057	0.0053	0.0053	0.0049
	(0.0036)	(0.0036)	(0.0036)	(0.0035)
I_{1011}	2.1184	2.2259	2.1882	2.3176
	(0.4134)	(0.4237)	(0.4177)	(0.4288)
I_{15}	1.8011	1.8673	1.8097	1.8913
	(0.3819)	(0.4001)	(0.3908)	(0.4063)
I_{16}	1.7248	1.8940	1.8352	1.9856
	(0.4007)	(0.4210)	(0.4129)	(0.4286)
Math pretest		-0.9701		-0.8986
		(0.1311)		(0.1348)
Game play			-0.6622	-0.5451
			(0.1154)	(0.1177)
Goodness of fit				
Deviance	628.505	564.318	591.873	541.144
AIC	642.505	580.318	607.873	559.144

Also, the coefficient of *Weeklygame* is -0.55. Its negative sign indicates that students who spend more time in video game play per week are less likely to fail in the game. As with the situation with prior math knowledge, students with more game experience were more successful in the game. The estimated odds ratio converted from the coefficient is 0.58, which means that each 1-point increase on the 4-point (the amount of) video game play scale is associated with a 42% decrease in the odds of failing in the game, controlling for other covariates. For example, for a student spending one to two hours per week on video game play, the estimated odds is 0.12, and for a student spending three to six hours per week on video game play and whose other characteristics are comparable, the estimated odds of failing is 0.07 (0.12 × 0.58).

Figure 5 displays the fitted hazard (the upper panel) and survival probabilities (the lower panel) from Model D. One can see the appreciable difference in the fitted hazards for different groups based on math pretest scores and the weekly amount of video game play. For students whose pretest scores are 1 *SD* below the average and who reported spending no hours per week in video game play, the risk of failing in the game is already fairly high in the early levels of the game. The estimated hazard probabilities are as high as around .4 and .5 in Levels 10 and 11 respectively. Thus, far more than half of the students fail in the game until Level 11. The risk is considerable in Levels 12, 13, and 14, where for other students the risk is rather minimal, and peaks in Levels 15 and 16. This means that among the already small number of students who are left to play the game, few students succeed in later levels in the game. Such a tendency is also shown in the estimated survival probability. The estimated median life time is around Level 10 for students whose pretest scores are 1 *SD* below and spend no hours in weekly video game play. Virtually no student is left to play above Level 15 in the game.

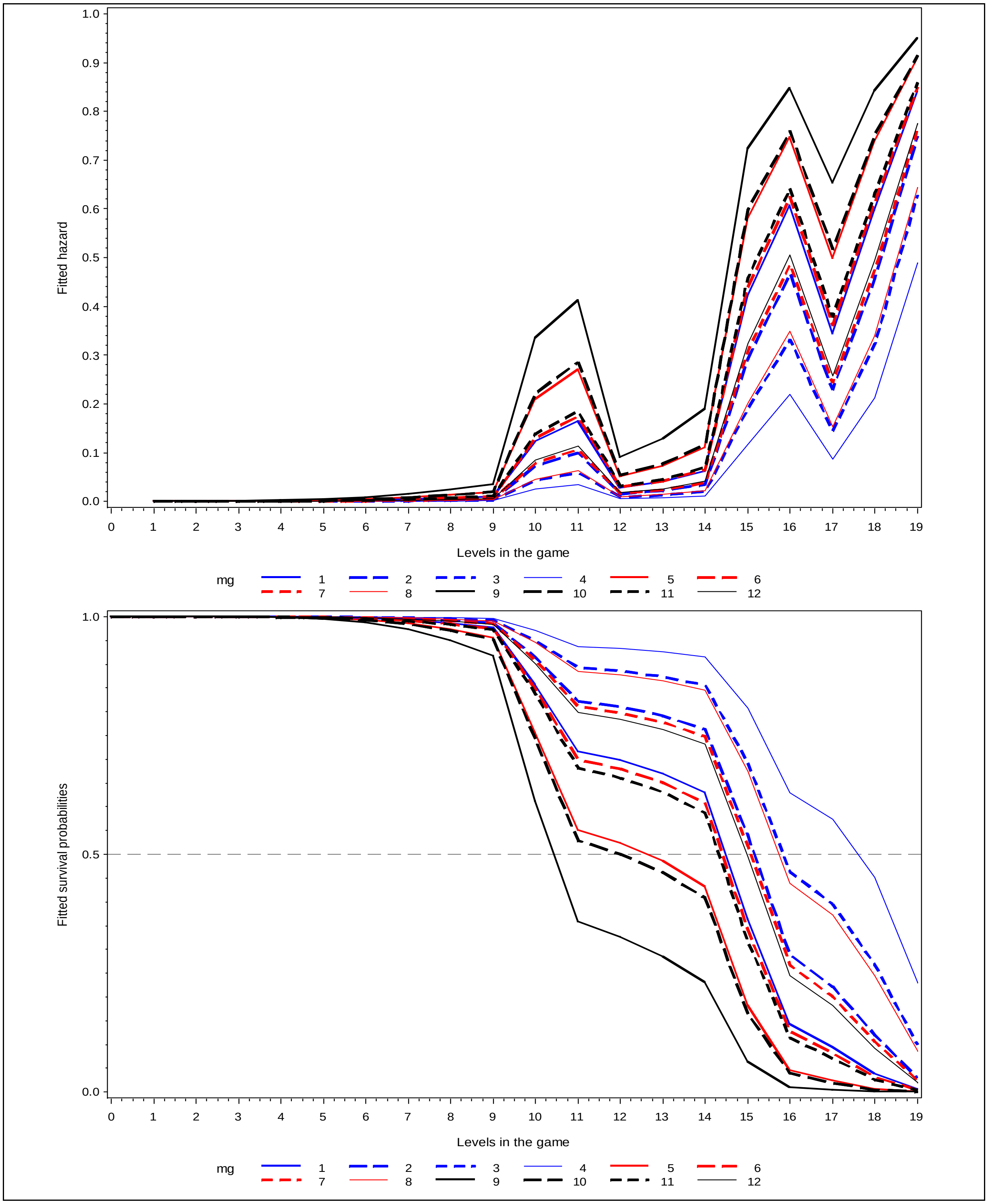


Figure 5. Fitted hazard probabilities (upper panel) and survival probabilities (lower panel) based on Model D (Equation 4) for different groups based on math pretest scores and the weekly amount of video game play.

For students whose pretest scores are 1 *SD* above the average and who have spent more than 6 hours per week in video game play, the risk of failing in the game is fairly minimal until later levels of the game: The estimated hazard probabilities are appreciably lower than .1 or close to 0 until the 14th level of the game. However, even for these students, the risk increases in Levels 15 and 16, and is fairly high in Levels 18 and 19. This means that a vast majority of these students are left to play the game in later levels in the game (e.g., 14 or above). As also shown in the plot of the fitted survival probabilities, the estimated median life time is around Level 18, or the second to last level of the game for these students. Thus, these students tend to succeed in advancing in the game until they fail in either the final level or the second to last level in the game.

Research Question 3: Do Different Game Versions Affect Student Performance?

Table 5 presents results from fitting a discrete-time hazard model as shown in Equation 5 to the data. The coefficient of *Gameversion* is -0.45, which is statistically significant (*p* value = -.05). Its negative sign indicates that students who played the experimental version of the game are less likely to fail in the game relative to students who played the baseline version of the game. In the odds scale, the estimated odds of failing to advance in the game is 64% of the estimated odds when students with similar characteristics play the experimental version of the game as compared to the baseline version of the game. For example, suppose that, for students who play the baseline version, the average odds of failing is 1 (hazard rate of .5). Then if students with similar characteristics play the experimental version of the game, the average odds of failing is only .64 (hazard rate of .39).

Table 5

Results of Fitting Discrete Time Hazard Models Comparing the Experimental and Baseline Versions of the Game

Variable	Coefficient	*SE*	Wald chi-square	*p* value	Odds
One	-10.5392	4.4417	5.6302	0.0177	<0.001
Time	1.6294	1.3167	1.5313	0.2159	5.10
Time^2	-0.1318	0.1207	1.1926	0.2748	0.88
Time^3	0.0047	0.0035	1.8385	0.1751	1.01
I_{1011}	2.3322	0.4292	29.5255	<.0001	10.30
I_{15}	1.8702	0.4059	21.2294	<.0001	6.49
I_{16}	1.9738	0.4291	21.1577	<.0001	7.20
Gameversion	-0.4461	0.2276	3.8396	0.0501	0.64
MathPretest	-0.9429	0.1378	46.8178	<.0001	0.39
Gameplay	-0.5441	0.1181	21.2142	<.0001	0.58

Conclusion and Discussion

Even very basic questions about game performance involves examining how game performance relates to other factors of interest. For example, why do some students successfully advance in the game, while others struggle and do not succeed at certain levels? Also, why in certain levels do some students show an abrupt increase in the risk of not advancing in the game? However, there has been no guidance, from a methodological standpoint, on rigorous ways to analyze and study game performance. Game performance is often measured by the farthest level a learner has reached in the game. Although the last game level a learner has reached during gameplay may intuitively appeal as a measure of game performance and may serve as a good measure for exploratory analysis, in settings where some learners possibly finish the final level in the game, naive analysis of the measure (i.e., the last game level learners have reached) may yield unbiased estimate of key parameters of interest.

This paper applies the DTSA technique to understand game performance and illustrates that it is a useful way to analyze game process data. DTSA enables studying the progression in the game of students with various characteristics, predicting heterogeneity across individuals in game performance over sequences of game levels. As noted in a number of places in this paper, in the presence of censoring (e.g., in this illustration, students who finish the final levels in the game, so students whose true level of game performance is unobserved), progression of students with various characteristics can be accurately estimated using the DTSA technique.

The concepts associated with DTSA, such as life tables and hazard and survival probabilities, provide a set of tools to show how learners proceed through a sequence of levels in the game. As illustrated in the game *Save Patch*, estimated hazard models can suggest overall trends of learners' hazard rates of advancing in the game across all game levels and at which levels learners tend to encounter more difficulties. These estimated DTHMs corresponded well with the hypothesis underlying the sequence of game levels in our study. When a level in the game introduces "big ideas" or higher loads of math knowledge, the hazard rates (probability of students failing to advance) at such levels abruptly increased. One of our primary interests was whether game performance data can provide a description of learning progression (Briggs et al., 2006; Wilson, 2009) of underlying math knowledge. This result may indicate that the tools that DTSA provides can be used as a systematic way for us to learn not only about the progression in the game but also about the learning progression of game players.

This possibility can also suggest the potential of using instructional games as diagnostic assessments. If DTSA applied to the game process data closely approximates the learning progression of game players, their stopping and struggling at a certain level can possibly show where learners' misunderstanding or lack of understanding comes from. If this piece of information (i.e., where, or at which level, game players failed to advance in the game) is combined with other information in the game process data (e.g., dwell time, frequencies and types of mistakes at certain levels), the accuracy of diagnostic information may drastically improve. This potential is supported by another finding in our study. Among a host of student characteristics that were measured before gameplay, the best single predictor of game performance was the measure of math knowledge prior to gameplay related to the math content of *Save Patch*. These ideas may naturally connect to ideas of formative assessment (Black et al., 2003; Black & Wiliam, 1998), if based on such information elicited from student behaviors, further instruction can also be planned either within the game or by instructors.

However, the potential of instructional games and of the applications of the DTSA technique should be tempered by possible caveats which also emerge from the study results. First, even though the potential of a game as a diagnostic assessment can be stimulating, it is hampered to some extent by findings in our study. While the best predictor of game performance was the measure of math knowledge prior to game play, the second best predictor of game performance was the weekly amount of time students spend on video game play. The DTHM fitted the data best when it includes both measures: previous math knowledge and previous video game experience. Thus, game performance appears to be a product of a combination of math knowledge and skills and game knowledge and skills. Although, in terms of averages, we control for one predictor in estimating relationships of the other variable to game performance, for each individual it is difficult to disentangle the two types of knowledge and skills (i.e., math and game) from their game performance. This suggests that although game performance can be used as a diagnostic or instruction-assisting tool, in the presence of empirical evidence indicating its validity, assessments based on game performance NEED TO BE VALIDATED BEFORE USED FOR ANY HIGH-STAKES DECISION. should not be used for any high-risk decisions.

Secondly, we also acknowledge a couple of caveats in using study results as "learning progression." Two main unresolved issues are: (a) how much agreement is sufficient between data and hypotheses based on sequences of levels in game development? and (b) would advancing in the game also be influenced by learning of game mechanics, or improvement in game skills? As for the agreement, in *Save Patch*, we do have very close approximation of data and the hypothesis. However, questions still arise. Our final model captured abrupt changes in Levels 10 and 11 and Levels 15 and 16 using binary indicators of these levels. By game design, Level 10 is where a big idea is introduced. Then why is there an abrupt change again in Level 11? In other words, why do students who pass Level 10 successfully still have a high risk in Level 11? The same question arises in Levels 15 and 16 too. One possible explanation may be due to the time limit (about 30 minutes)—for example, students who passed Level 10 failed to pass Level 11 since time was up for the administration—but this will need to be confirmed in a future study. A skeptic can legitimately ask whether advancing in the game may be due to improvement in game skills, which is possible, especially for students who came with minimal experience with playing video games. Thus, "learning" progression may not be learning the math content of the game such as knowledge of fractions in *Save Patch*, but rather an educationally meaningless process of learning game mechanics. Skepticisms of this kind can be best refuted when there is good agreement between models based on the data and the hypothesis based on the game design. Especially when there are abrupt changes in hazard rates exactly where mathematically "big ideas" come in, and when these levels do not especially require additional game skills than previous levels, the progression through the game levels will most likely be the progression of the learning of the math content.

The main purposes of math instructional games can differ in each game, and depending on the purposes, they should be designed in different ways. This study presents some suggestive findings from a particular game, *Save Patch*, about the possibilities for instructional math games to be used as an assessment of student learning progression. If the purpose of a game would be such assessments that can assist diagnostic assessments, instructional planning, or more generally ideas of formative assessments (Heritage, Kim, Vendlinski, & Herman, 2009; Sadler, 1989; Schifter, 1998), further validity studies are strongly warranted. Learning from the study in this paper, we point out the following. First, the main validity study needs a pretest, of which the content is aligned with the content of the game: a pretest that consists of items developed to measure the sequence of math content in the game of interest. Second, the application of DTSA as illustrated in this paper provides a general framework for such validity studies and allows us to draw accurate inferences concerning game performance in relation to various factors of interest. Third, the main confounding factors in studying learning progression were previous video game experience, and the time limit in the administration setting. Thus, future validity studies should be designed in advance to minimize such confounding factors. In sum, learning from this study, some possible ways to reduce the confounders include: (a) to design games that would not involve high levels of game skills and that would need constant, not increasing, levels of game skills as game levels advance; (b) to expose students to such gameplay in advance so that student game performance would not be mediated considerably by game skills; (c) or alternatively, for validity studies, to select students with some previous exposure to video games and some game skills as a study sample; and (d) to give sufficient time for all students to try to finish the game unless they do not know how to solve the math in the game.

Lastly, we would like to note that another potential of using survival analysis techniques in game performance data comes from the capability of incorporating time-varying predictors. DTSA readily provides a way of examining the relationships between time-varying predictors and game performance. This may prove powerful in accommodating game-process data in the analysis such as dwell time, frequencies, and types of mistakes at a certain level, since such game process data are considered as time-varying covariates in the survival analysis framework. It is notable that game process data will be more and more available, since current technologies allow us to obtain clickstream data from each individual who played a game.

In addition, various study designs can be made possible by incorporating time-varying covariates. This can help test various hypotheses and draw sound inferences between the time-varying covariates and game performance. For example, as students learn more about curricular units that are critical in understanding the domain, their performance in the game may change. Also, instructions that use different methods (e.g., ones that are sequenced in different orders) may result in differential patterns of risks of experiencing events over levels, as well as potentially yielding higher or lower overall risks. Relevant instruction that happened in the middle of gameplay may also be related to the performance in the game.

References

Allison, P. D. (1995). *Survival analysis using the SAS system: A practical guide.* Cary, NC: SAS Institute Inc.

Baker, E. L., Chung, G. K. W. K., & Stigler, J. (2010). *IES Annual Report: Year 2 Center for Advanced Technology in Schools (CATS)* (Deliverable to IES). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. New York, NY: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–73.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple choice items. *Educational Assessment, 11*(1), 33–63.

Center for Advanced Technology in Schools. (2010). *Save Patch game level design: RN1.2 (July/November 2009).* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association, 83,* 414-425.

Fahrmeir, L., & Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association, 91,* 1584-1594.

Federation of American Scientists (FAS). (2006). Summit on educational games: Harnessing the power of video games for learning. Washington, DC: Author.

Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion* (Tech. Rep. No. 2005-004). Orlando, FL: Naval Air Warfare Center Training Systems Division.

Heritage, M., Kim, J., Vendlinski, T. P., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice, 28*(3), 24-31.

Hosmer, D. W., Jr., & Lemeshow, S. (1999). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144.

Schifter, D. (1998). Learning mathematics for teaching: From the teacher's seminar to the classroom. *Journal for Mathematics Teacher Education, 1,* 55-87.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis.* New York, NY: Oxford University Press.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64,* 489–528.

Tobias, S., & Fletcher, J. D. (Eds.). (2011). *Computer games and instruction*. Charlotte, NC: Information Age Publishers.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716-730.

Appendix

Table A1

Goals and Knowledge Specifications in *Save Patch* Game Levels

Level	Goal(s)	Targeted knowledge specification
1	Place 1 trampoline on block to move 1 space. Space = 1 unit	1.1.0. The size of a rational number is relative to how one whole unit is defined. 1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.). 1.3.0 In our number system, the unit can be represented as one whole interval on a number line.
2	Place 1 trampoline on block to move 2 spaces. Space = 1 unit	1.1.0. The size of a rational number is relative to how one whole unit is defined. 1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.). 1.3.0 In our number system, the unit can be represented as one whole interval on a number line. - 1.3.1. Positive integers are represented by successive whole intervals on the positive side of zero. - 1.3.2. The interval between each integer is constant once it is established. - 1.3.4. All rational numbers can be represented as additions of integers or fractions. 2.1.0 To add quantities, the unit (or part of unit) quantities must be identical. - 2.1.1. Identical (or common) units can be descriptive (e.g. apples, oranges, and fruit) or they can be quantitative (e.g. identical lengths, identical areas, etc.). - 2.1.2 Positive integers can be broken into parts that are each one unit in quantity. These identical units can be added to create a single integer sum. 2.2.0. Identical (common) units can be added to create a single numerical sum.
3	Place trampoline(s) on blocks to move 2 spaces Space = 1 unit	Same as Level 1.
4	Place trampoline(s) on blocks to move 4 spaces Space = 1 unit	Same as Level 2.

Level	Goal(s)	Targeted knowledge specification
5	Place trampoline to move 2 spaces Space = 1/2 unit Move 1 whole unit but on this level unit is in two pieces.	1.1.0. The size of a rational number is relative to how one whole unit is defined. 1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.). 1.3.0 In our number system, the unit can be represented as one whole interval on a number line. - 1.3.3. Positive non-integers are represented by fractional parts of the interval between whole numbers. - 1.3.4. All rational numbers can be represented as additions of integers or fractions. 2.1.0 To add quantities, the unit (or part of unit) quantities must be identical. - 2.1.1. Identical (or common) units can be descriptive (e.g. apples, oranges, and fruit) or they can be quantitative (e.g. identical lengths, identical areas, etc.). 2.2.0. Identical (common) units can be added to create a single numerical sum. 3.1.0. The denominator of a fraction represents the number of identical parts in one whole unit. That is, if we break the one whole unit into “x” pieces, each piece will be “1/x” of the one whole unit. 4.1.0. The numerator of a fraction represents the number of identical parts that have been combined. For example, ¾ means three pieces that are each ¼ of one whole unit. 5.3.0. Any rational number can be written as a fraction that relates one integer—the number of parts there are (numerator)—to another integer—the number of parts in one whole (denominator).
6	Move three spaces Space = 1/2 unit Convert so all the coils are the same size.	Same as Level 5 and: - 2.1.3. Each Whole Unit or part of a Whole Unit (fractions) can be further broken into smaller, identical parts, if necessary. 2.3.0. Dissimilar quantities can be represented as an expression or using some other characterization, but are not typically expressed as a single sum [NB: we are considering numbers like 2 ¾ to have an implied addition – so 2 + ¾ – whereas 11/4 is a single sum].

Level	Goal(s)	Targeted knowledge specification
7	Move 4 spaces Space = 1/2 unit Convert the coils to 1/2 and add the coils to the trampolines.	1.1.0. The size of a rational number is relative to how one whole unit is defined. 1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.). 1.3.0 In our number system, the unit can be represented as one whole interval on a number line. - 1.3.1. Positive integers are represented by successive whole intervals on the positive side of zero. - 1.3.2. The interval between each integer is constant once it is established. - 1.3.3. Positive non-integers are represented by fractional parts of the interval between whole numbers. - 1.3.4. All rational numbers can be represented as additions of integers or fractions. 2.1.0 To add quantities, the unit (or part of unit) quantities must be identical. - 2.1.1. Identical (or common) units can be descriptive (e.g. apples, oranges, and fruit) or they can be quantitative (e.g. identical lengths, identical areas, etc.). - 2.1.2. Positive integers can be broken (decomposed) into parts that are each one unit in quantity. These single (identical) units can be added to create a single numerical sum. - 2.1.3. Each whole unit or part of a whole unit (fractions) can be further broken into smaller, identical parts, if necessary. 2.2.0. Identical (common) units can be added to create a single numerical sum. 2.3.0. Dissimilar quantities can be represented as an expression or using some other characterization, but are not typically expressed as a single sum [NB: we are considering numbers like 2 ¾ to have an implied addition – so 2 + ¾ – whereas 11/4 is a single sum]. 3.1.0. The denominator of a fraction represents the number of identical parts in one whole unit. That is, if we break the one whole unit into “x” pieces, each piece will be “1/x” of the one whole unit. 4.1.0. The numerator of a fraction represents the number of identical parts that have been combined. For example, ¾ means three pieces that are each ¼ of one whole unit. 5.3.0. Any rational number can be written as a fraction that relates one integer—the number of parts there are (numerator)—to another integer—the number of parts in one whole (denominator).
8	Move 5 spaces Space = 1/3 unit	Same as Level 5.
9	Move 6 spaces Space = 1/6 unit	Same as Level 7.
10	Move 6 spaces Space = 1/6 unit Find lowest common denominator.	Same as Level 7.

Level	Goal(s)	Targeted knowledge specification
11	Move 5 spaces Space = 1/3 unit Find lowest common denominator.	Same as Level 7.
12	Move 2 spaces Space = one unit Direction matters.	Same as Level 1.
13	Move 4 spaces Space = one unit Direction matters.	Same as Level 1.
14	Move 4 spaces Space = 1/2 unit Direction matters. Break whole coil into 1/2.	1.1.0. The size of a rational number is relative to how one whole unit is defined. 1.2.0. In mathematics, one unit is understood to be one of some quantity (intervals, areas, volumes, etc.). 1.3.0 In our number system, the unit can be represented as one whole interval on a number line. - 1.3.3. Positive non-integers are represented by fractional parts of the interval between whole numbers. 3.1.0. The denominator of a fraction represents the number of identical parts in one whole unit. That is, if we break the one whole unit into “x” pieces, each piece will be “1/x” of the one whole unit. 5.3.0. Any rational number can be written as a fraction that relates one integer—the number of parts there are (numerator)—to another integer—the number of parts in one whole (denominator).
15	Move 18 spaces Space = 1/3 unit Direction matters Need to figure out path. Break whole coils into 1/3.	Same as Level 7.
16	Move 10 spaces Space = 1/2 unit Direction matters. Need to figure out path. Find lowest common denominator.	Same as Level 7.

Level	Goal(s)	Targeted knowledge specification
17	Move 10 spaces Space = 1/2 unit Direction matters. Need to figure out path. Find lowest common denominator.	Same as Level 7.
18	SAME AS LEVEL 15	Same as Level 7.
19	Move 22 spaces Space = 1/3 unit Direction matters. Need to figure out path. Find lowest common denominator.	Same as Level 7.