Choice-based Assessment: Can Choices Made in Digital Games Predict 6th-Grade Students' Math Test Scores?

Min Chi
Department of Computer Science
North Carolina State University
Raleigh, North Carolina 27695
mchi@ncsu.edu

Daniel L. Schwartz
AAA Lab
Stanford University
Stanford, California, 94305 USA
danls@stanford.edu

Kristen Pilner Blair
AAA Lab
Stanford University
Stanford, California, 94305 USA
kpilner@stanford.edu

Doris B. Chin
AAA Lab
Stanford University
Stanford, California, 94305 USA
dbchin@stanford.edu

ABSTRACT

In this paper, we mined students' sequential behaviors from an instructional game for color mixing called Lightlet. Students pkaying the game have two broad strategies. They can either test candidate color combinations in an experiment room without risking an incorrect answer. Or they can choose colors from a faux shopping Catalog containing several different mixing charts. While the results shown in the Experiment Room are always correct, only a few of the charts in the Catalog are correct. Thus, if students use the catalog students must apply critical thinking skills to determine what charts to trust. Our primary goal in this work was to identify the crucial choice pattern(s) in students' game play that would contribute to their learning or subsequent performance. Data was collected from 6th graders. The results showed that children who chose to explore the Catalog of different charts during the game performed better in school. More specifically, the types of behavior choices students committed during the game play predicted about 43% of the variation in their subsequent math grades. This project shows that by assessing students' choices during learning, we can discover a great deal about their learning process and can identify and assess choices that are critical for learning but are often missed by most tests.

Keywords

Educational Assessments, Educational Games, Choice-based Assessment, Mining Behavior Data

1. INTRODUCTION

Educational assessment sits at the epicenter of learning research. Any quantitative study of an intervention, experience, or program to improve learning depends on the quality of the outcome measures. An ideal educational assessment would both reflect and reinforce the educational goals that society deems valuable. One fundamental goal of education is to prepare students to act independently in the world—which is to make good choices. It follows that an ideal assessment would measure how well we are preparing students to do so.

Most existing educational assessments are knowledge-based in that they focus on the amount of knowledge and skills students have accrued. Many such assessments use a format that Bransford and Schwartz (1999) labeled, Sequestered Problem Solving (SPS). In the typical SPS assessment, students are sequestered from learning opportunities and outside resources that might contaminate the validity of the assessment. Bransford and Schwartz argue that these retrospective measures are appropriate if the goal of instruction is training for highly stable performance conditions, but they are not optimally diagnostic when the goal is to prepare students to continue adapting and learning.

As an alternative, Bransford and Schwartz followed the theories of Vygotsky (1934) and Fueurstein (1979) to propose Preparation for Future Learning (PFL) assessments. In a PFL assessment, there are opportunities for learning during the assessment process, and the question is whether students are prepared to take advantage of these opportunities. These types of assessments are appropriate when the assumption is that students will need to continue learning, and the question is whether prior instruction and experiences have prepared them to do so. Multiple studies have shown the value of including PFL measures for assessing the quality of classroom instruction (Schwartz & Bransford, 1998; Schwartz & Martin, 2004; Chin, et al, 2010).

In the present work, we brought PFL assessments into an interactive context, where it is possible to directly measure processes associated with PFL. In our approach, choices, rather than knowledge, was the interpretative frame within which learning assessments are organized. In the following, we refer our approach as choice-based assessment.

Until recently, most researchers treated choice as a form of learning intervention. Iyengar and Lepper, (1999), for example, argued that giving students choices can increase their motivation and learning. Choice is important for learning if only because students need to experience choices in the protected atmosphere of education so they can learn how to handle them before becoming independent. Our approach is different. We ask why choices should be viewed as the outcome of learning and not solely an instructional ingredient to improve it. We contend that choice should be the interpretative framework for understanding and assessing learning outcomes. With new developments in technology, it should be possible to advance this goal which was beyond the reach of prior assessments.

One particularly promising way to integrate choice into educational assessment involves creating process measures that can capture student behaviors dynamically. Digital technologies make it possible to teach and assess student learning in new ways. Simply put, many new technologies are about choice. When browsing webpages, each click can be considered a choice about learning. When deciding what online sources to trust and which friends to consult, people are making learning choices. When using scientific simulations, people choose which sequence of

settings that will yield the most telling results. Thus there is a good match between digital technologies and choice-based assessments.

Digital technologies make choice-based assessments possible, because interactive assessments can evaluate students in the context of choosing whether, what, how, and when to learn. By logging what students choose to do in an interactive environment, it is possible to gather functional process measures that can be expensive and difficult to gather by other means (e.g., Aleven et al., 2003; Baker, Corbett, & Koedinger, 2004; Hogyeong et al, 2008; Stevens & Thadani, 2007). Cognitive Tutors (Koedinger & Anderson, 1997), for example, track student progress and actions across multiple hours of use. Videogames include metrics of success and process (Gee, 2003).

However, the examples, which we describe below, all depend on large-scale environments that require many hours of interaction before any useful information can be gathered. To serve a broad range of goals, assessments need to be more nimble. We show that this is possible by demonstrating a digital choice-based assessment designed to assess critical thinking. This assessment is drawn from our work on Choicelets. It may not be what you expect in a test.

There are advantages to making smaller and more nimble environments for assessment. First, nimble assessments do not depend on students completing many hours of a complex game or instructional sequence before it is possible to make any useful assessments.

Second, smaller assessments can target specific choices design. This is quite different from searching for diagnostic patterns amid the millions of possible choice combinations in larger open environments.

Third, there is value to having an assessment that can be used to make general comparisons. In video games, cognitive tutors, and many embedded assessment tools, the assessments are locked into a specific model of instruction and delivery system. Thus they cannot be used to compare the effectiveness of different instructional models and learning experiences.

2. Choicelet

Choicelets take the form of short, and hopefully, engaging games that students want to complete. To complete the game, each Choicelet requires some learning, and we keep a log of students' choices during the process. Different Choicelets are designed to assess specific constellations of choices relevant to learning. In the present work we will focus on Lightlet a game designed to assess students' critical thinking skills. Most assessments of critical thinking evaluate deductive reasoning; for example, the ability to recognize when assumptions do not lead to conclusions. We chose instead to reclaim the broader meaning of critical thinking as the process of rationally deciding what to believe (Norris, 1985). Therefore, in Lightlet, we assess the decision to engage in critical thinking for the purpose of learning.

Figure 1 shows the main interface of the Lightlet. To play Lightlet, students mix the primary colors of light to move through a series of puzzle levels. The main component is a game board with colored light tiles, shown in the center of the screen. The first step in the game is to pick a colored tile from the game board. This is the *target color*. There is no constraint on the ordering of tiles to play and students can play them in any order. The students then select from the colors shown below the game board. The colors include the three primary colors (red, green, blue) and one non-primary color. With a click of the mix button, the selected colors mix. If the mixed colors produce the target color, then the tile disappears from the game board and reveals a portion of a rebus (picture that makes a phrase); otherwise, the tile remains. There is no limit on the number of times that a student can try for each tile. Once students remove all of the tiles from the board or reveal enough of the rebus to guess it correctly, then they can move to the next level of the game.

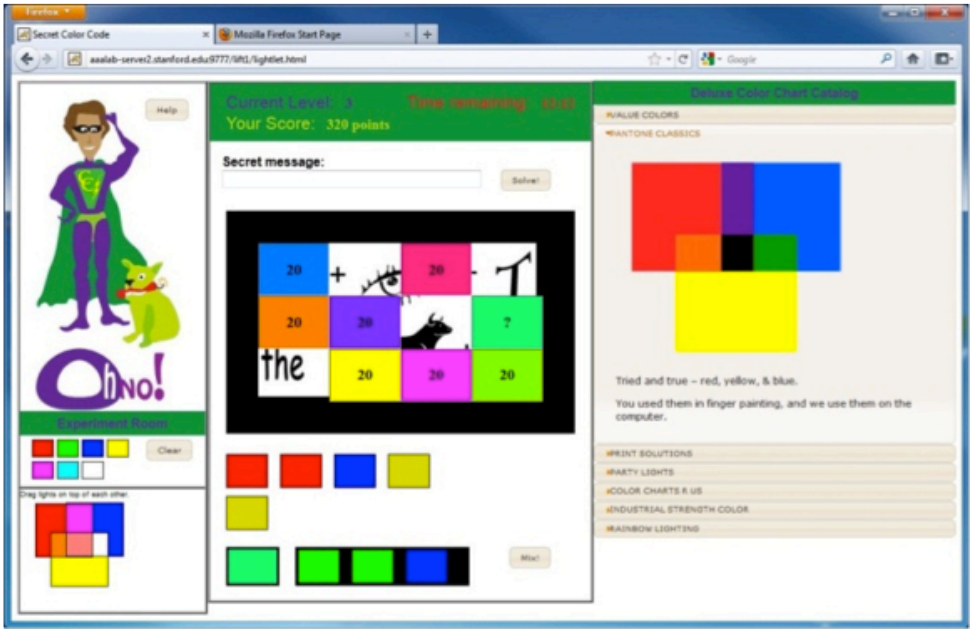

Figure 1. Lightlet GUI showing the Experiment Room (lower-left), Game Board (center) and Catalog (right).

Most students know about the primary colors for mixing paint or *subtractive color*. They are: red, yellow and blue. Mixing light or *additive color* however, depends upon a different set: red, green, and blue (RGB). Red + green makes yellow, and red + green + red makes orange. Thus, a major part of the game involves learning about additive color.

Lightlet includes a pair of resources to help students learn both of which are shown in Figure 1. On the lower left-hand side of the screen is the Experiment Room, where students can try out different color combinations without risking a wrong answer in the game board. In the Experiment Room, students can use seven base colors including the three primary colors. There is no upper limit on the number of times that each base color can be used. Students can clear the experiment room at any time with the 'clear' button.

On the right, there is a faux shopping Catalog in which different companies sell charts for mixing colors. There are seven charts available. Two of them are for additive color (i.e. light) and are correct. The remainder are for subtractive color mixing (i.e. paint) and are thus incorrect. The chart shown in Figure 1 is incorrect and is designed to play into students' prior beliefs about mixing paint. The descriptive text for the chart says: "Tried and true, red, yellow, and blue. You used them in finger painting! Use them now." The Experiment Room correctly shows that yellow and blue make white while the Catalog entry shows that they make green. Students must use critical thinking to decide which charts to believe, if they choose to use them at all. We track each of the students' choices during the game play in the log files.

There are three levels in Lightlet. The introduction (level 1), simply involves mixing red and blue lights to make three colors: red, blue, and magenta. This task is easy as it conforms to students' existing assumptions about mixing paint. Only the experiment room is available in this level.

Level 2 provides for full game play. In this level the color mixing challenges become harder in that students can use all three primary colors (e.g. red + green = yellow); we open both resources (the Catalog and Experiment Room) so that players can use them to figure out how to mix light. It is in this level where we collect the process data of most interest. Students may be induced to use *trial and error* in the Experiment Room or using the chart Catalog. If the students who used the chart Catalog did better, then there would be a warrant that this is a better choice pattern.

In the more advanced level (level 3) students are provided with new, even harder challenges as students can use each primary color twice (e.g., making orange, which requires red+red+green). And both the Catalog and Experiment Room are available for all players. Figure 1 shows this level of the game, which includes color combinations that depend on mixing three lights together.

3. Two Dominant Choices on Lightlet

There were two dominant patterns of choices in Lightlet: One pattern, The catalog-related choice pattern occurred when students chose to figure out which of the color charts are correct.

The other pattern, the Experiment Room-related, happens when students use the Experiment Room to solve the problems. These students would mix colors in the Experiment room to determine which colors to mix for the gameplay. Once they find the answer in the Experiment Room, they then choose the corresponding color on the game board and mix colors correctly.

In brief, on Lightlet, students have to learn the rules of additive color, they have an Experiment Room in the lower left corner, and they have a set of Catalog charts that show different color mixing results, some of which are subtractive charts and some of which are additive. Our question is: Do students choose to engage in critical thinking by deciding what charts to believe or they choose to try-and error in Experiment Room?

We hypothesize that students who applied the Experiment Room choice pattern had learned to solve problems one at a time rather than trying to find a general explanation. In math class, one can imagine students working to get the right answer for each separate math problem without attempting to find the deeper explanation that handles all possible related problems. The students who spent their time trying to decide which color chart to believe, on the other hand, were trying to find the *general* framework that could handle any colors in the game.

4. Teaching Students General Explanation

Prior research has shown the advantage of asking students to generate a general explanation that can handle all cases in a given task. It is much like finding a good theory can explain the results of multiple experimental conditions. For example, in a series of studies, students were provided with sets of contrasting cases designed to help them induce the structure of density (Schwartz, et al. 2011). When students were asked to invent a single procedure for generating a "crowdedness index" for the cases, they learned the ratio structure of density and spontaneously transferred to new problems. They outperformed control students, who were told about density at the outset and then applied the formula to the exact same contrasting cases.

In the current study, we first conducted pre-Lightlet training on generating general explanation. During the training, students were asked to find the similarities and differences between a series of contrasting cases in two physics tasks for two consecutive Friday classes (50 min each), one task per class. Half of the students, the experimental group, were explicitly encouraged to produce a *comprehensive, general* explanation of the similarities and differences. In other words, the experimental group was tasked with finding an underlying general structure or framework that explains all contrasting cases while the control group were tasked with finding the similarities and differences between the cases and they were not explicitly tasked to generate any general explanations or framework.

The two groups were treated identically when using Lightlet. Because the task of determining which set of charts to trust would help students find a *general* framework that could handle any colors in the game, we hypothesize that the experimental students in the pre-Lightlet training would be more likely to do so, especially since they were not only explicitly taught to do so but also greatly benefited from generating a general explanation for both physics tasks. So our first hypothesis is that: *the experimental students will be more likely to use Catalog charts than the control students.*

5. Validating Choice-based Assessments

In the normative world of education, we care about "better" and "worse," so it is crucial to justify whether or not a particular performance is "better." Knowledge-based assessments rely on objective "right" and "wrong" answers as their criteria for better and worse performance. Few would argue with the claim that "five" is a worse answer than "four" to the question, "What is two plus two?" But with choices, people may reasonably challenge the

judgment that one choice is better than another. Who is to say that persisting is better than not? If we were to analyze the log file of a student using Lightlet, for example, the choice of where to let the cursor rest while thinking is less relevant than the choice of whether to open up one of the color charts. A data-driven answer would help alleviate some of the problems associated with the social construction of what constitutes a useful choice, at least with respect to learning. People would then be able to debate whether the learning value of a choice is sufficiently high to favor it in assessment.

Here we present some approaches to validating whether some choices are better or worse than others. As always, our criteria of better and worse are relative to learning. We begin with a correlational approach: whether certain choices are connected to standard knowledge-based measures. This is important since most educators still think knowledge-based assessments are the ground truth for evaluating learning. In the following, we used the students' final school math test scores as a *standard* knowledge-based assessment. More specifically, we will investigate whether the choices students made when interacting with Lightlet would predict their final school math test scores.

To further validate the choice-based assessment, we will directly compare the choice-based assessment with *game-embedded* knowledge-based assessments such as how well did a student play Lightlet. Thus our general question is: which assessment is predictive to students' school performance, the choice-based assessment, the *embedded* knowledge-based assessment, or neither. Given that Lightlet has little to do with solving math problems as they appear on the children's mathematics tests, we hypothesize that that *game-embedded* knowledge-based may not be very predictive.

As described above: we hypothesized that students involved in Catalog-related choice patterns are interested in finding a *general* framework that could handle any colors in the game. Since the Catalogs only became available on the full game play level (level 2), we expect general-explanation students would begin exploring the Catalog choices on level 2, immediately or shortly after each catalog becomes available. Therefore, our second hypothesis is that: *when considering level 2 alone the choice-based assessment will be a better predictor of students' math performance than the game-embedded knowledge-based assessment.*

On the other hand, considering level 2 alone may not be sufficient to grasp all the students' choice patterns. On average, students spent only around 5 minutes on level 2 vs. 20 minutes on the whole game. So the effectiveness of choice-based assessment may become even more predictive if we considering the entire game play. Therefore, our third hypothesis is that *when considering the whole game play, the choice-based assessment will still be a better predictor for students' math performance than the game-embedded knowledge-based assessments.* However, it is not clear whether *the former will be more effective than the choice-based assessment when using level 2 alone.* In other words, whether the longer the choice-based assessment is, the more effective it will be?

6. Data Collection

6.1 Participants and Design:

Two 6th-grade classes participated in the study. Both classes were from high-SES schools in California and had the same math teacher. Due to logistical constraints, intact classes were randomly assigned to the two conditions during pre-Lightlet training. The assignments were: Experimental (n = 19) and Control (n = 21). In both conditions, students variously worked individually or in groups, consistent with regular classroom practice. Then all students played Lightlet for 15-30 minutes. All tests in the study were taken individually.

6.2 Procedure:

The study occurred on three Friday classes (50 min each) with two consecutive Fridays for the pre-Lightlet training and one for interactions with Lightlet.

During the pre-Lightlet training, all students were given a set of contrasting cases on "cannon rides" shooting straight out at 0° angle at different speeds and from different heights in the first week and cases on "cannon rides" shooting out at an angle at different speeds and reaches different maximum heights in the second week. The treatment difference occurred in the instructions that students received. Control students were prompted to explain the similarities and differences among the cases while the experimental ones were told to invent a single general framework that would explain all cases. This phase lasted 15 minutes. Students answered a brief test item, used the simulation to test their ideas, and then took the posttest.

Interactions with Lightlet happened six weeks after pre-Lightlet training. All students played Lightlet for fifteen to thirty minutes. The students' final school math test was taken at the end of the semester, about one month after interacting the Lightlet.

6.3 Data Features

In order to identify students' choices and track their game-embedded performance during interactions with Lightlet, we defined a set of Catalog-related, Experiment Room related, and performance-related features based on a combination of theory and prior work on modeling learning environments (Chi, VanLehn, Litman, & Jordan, 2011) and on student modeling (Chi, Koedinger, Gordon, Jordan, & VanLehn, 2011). We do not yet know precisely what choice or performance actions are associated with learning outcomes in advance. Thus, we defined four categories of features.

The first two categories correspond to the two types of the choice patterns in section 3. More specifically:

The first category includes 18 features related with the Catalog usage. It includes two types of features: duration and occurrence. The former covers 11 time-related features such as the total duration that a correct or incorrect Catalog is open; while the latter includes seven features such as the total number of times that a student opened an incorrect Catalog, the total number of Catalogs that the student opened and so on.

The second category includes 13 features related with usage of the Experiment Room. It consists of one feature covering how much time a student spent in the Experiment Room and 12 features related to their behaviors within the room behaviors. These are simple features such as the number of times the students used the Experiment Room and more complicated features such as the number of times that a student successfully makes a target color in the experiment room after picking it on the game board.

The third category includes 6 features focused on general information about the game play. These include simple features such as the total time spent in the game, gender and the condition information during the Pre-Lightlet training. Some more complicated features in this category assess how students choose the next puzzle to play. On Lightlet, students were given 6, 9 and 12 puzzles on level 1, 2 and 3 respectively. For each level,

students can select which tile to play in any order. We thus defined three features to detect how the students choose the target. We found that, rather than using the Experiment Room, students sometimes engage in trial and error on the game board. One example feature is TryErrorPick, which is defined as the number of times a student picked a color from the game board that they had previously made a wrong attempt on. For example, if a student trying to mix "orange" by mixing red + green, the student would get yellow and the orange puzzle remains unsolved on the game board; if the student then chose yellow again this would be detected.

The last category includes 13 features related with game-embedded performance-related features. They include features such as the total number of tiles that a student succeeded, the percentage of tiles that a student succeeded (how well a student clear the game board), how efficient a student was when clearing the tiles (the number of tiles students cleared from game board divided by the total time) and so on.

7. Results

In the following, we will present our results in the order of our three hypotheses:

Hypothesis 1: *the experimental students will be more likely to use Catalog charts than the control students.*

Hypothesis 2: *when considering level 2 alone the choice-based assessment will be a better predictor for students' math test school than the game-embedded knowledge-based assessment.*

Hypothesis 3: *when considering the whole game play, the choice-based assessment still be a better predictor than the game-embedded knowledge-based assessments.*

7.1 Experiemental vs. Control

Our overall results show that the experimental and control groups were comparable at the outset of Pre-Lightlet training when the treatment differences began: there were no significant differences between treatment groups on two tests given by the teacher before the study, or any of our assessments before or at pretest on the first week of Pre-Lightlet training. As expected, after the different treatments, the two groups began to separate. We found that explicitly asking students to generate a general explanation led the experimental group to outperform the control condition on several midtest and posttest items after the different treatments took place. More specifically, 100% and 77.8% of experimental group produced a general explanation for the two physics tasks respectively while only 10% and 33.3% of the control group did so. This difference was statistically-significant for the first task and marginally-significant for the second.

The experimental group significantly outperformed the control group. We argue that this is because the former were asked explicitly to generate a general explanation for all the cases. On Lightlet, to generate a general framework for all the color games would requires students to engage in Catalog activities. We would expect that the experimental group would be more likely to do so. However, our results showed that the experimental students were no more likely to engage in Catalog activities than the control students. We compared the two groups across all Catalog -related features described in previous section on both level 2 and across the whole game. No significant difference was found on any of Catalog-related choices students made.

Furthermore, we found no significant difference between the two conditions on any Lightlet Experiment Room-related behaviors, game-board performance related, or the school math final test scores.

Overall, it seems that after explicit instruction on generating general explanations, the experimental students did not spontaneously make a choice that would lead to finding a general solution for all the colors on Lightlet. There are many possible explanations for this finding. One possible explanation is that the instruction on generating a general explanation was explicit during the Pre-Lightlet training but when interacting with Lightlet, the experimental students were not explicitly asked to do so. Additionally, the experimental students were given a set of comparing and contrasting worked cases in the original general explanation instruction; but on Lightlet, they were not given any.

To summarize, our results showed that it is still an open question how to teach students to make good choices. On the other hand, while it may not be easy to teach students to make good choices, is it still feasible to use choices as an effective assessment? In the following, we investigated whether the choices students made on Lightlet would predict their learning performance in school. We first investigated whether using level 2 data alone would be predictive.

7.2 Choice- vs. Knowledge-based Assessment Using Level 2 Only

We first investigated whether the individual features from the four categories would predict the standard knowledge-based assessment: students' final math test scores. Note that, all the features were calculated based on the level 2's log files alone. Among the four categories, Out of 18 Catalog-related features, 13 features are significantly predicted students' final math test scores. For all 13 features, the more Catalog activities, the higher the students' final math tests scores. Among them, the most predictive feature is: *ResourceReviewDuration* the total duration that a Catalog is open in level 2. *ResourceReviewDuration* significantly predicted students' final math tests scores: $\beta = 0.010$, $t(38) = 2.64$, $p = 0.01$. It alone also explained a significant proportion of variance in students' final math tests scores, $R^2 = .15$, $F(1, 38) = 6.95$, $p < .001$.

Only one out of 13 Experiment-Room related features are significantly predicted students' final math test scores. The feature is GoalOrientedExperimentSuccessTry: the number of times students pick a color from the game board and then try to make the targeted color successfully in the Experiment Room in level 2. GoalOrientedExperimentSuccessTry significantly predicted students' final math tests scores: $\beta = -1.93$, $t(38) = -2.69$, $p = 0.01$. It alone explained a significant proportion of variance in students' final math tests scores, $R^2 = .16$, $F(1, 38) = 7.22$, $p = .01$. So the more a student try and error successfully in the experiment room, the lower his/her final math score is.

Finally, none of the features in the remaining two categories, the general information and the game-embedded performance related features significantly predict students' final math test scores. For example, PercCorrectGamePlay: the total number of times tile a student succeeded divided by the total number of tiles the student tried to play in level 2, did not significantly predict student's final math scores: $R^2 = .05$, $F(1, 38) = 1.85$, $p = .18$.

Therefore, when considering level 2 alone the choice-based assessment is a better predictor for students' math test school than the game-embedded knowledge-based assessment when using the single feature.

We then applied brute-force search to select the best three features from all the four categories that would best predict students' final math test scores. Our final model include three features and they are:

> WrongResourceReviewDuration (Catalog-related): The total duration of a student opening a wrong Catalog
>
> GoalOrientedExperimentSuccessTry (Experiment Room related): The number of times students pick a color from the game board and then try to make the targeted color successfully in the Experiment Room.
>
> TryErrorPick (General): The number of times students pick a color from the game board that is the same color as the previous wrong color.

The results of the regression indicated the three predictors explained a significant proportion of variance in students' final math tests scores: R^2 = .43, $F(3, 36) = 8.94$, p = .0001. Table 1 shows that all three features significantly predict students' final math tests scores.

Table 1: Coefficient of Three Level 2 Feature Model

Feature Name	β	Sig
WrongResourceReviewDuration	0.018	$t(38) = 3.51$ p =0.001
GoalOrientedExperimentSuccessTry	-2.62	$t(38) = -4.14$ p =0.0002
TryErrorPick	-3.28	$t(38) = -2.69$ p <0.05

Finally, note that none of the game-play performance related features were included in the final three-feature model. Therefore, it again suggested that the choice-based assessment is a better predictor for students' math test school than the game-embedded knowledge-based assessment.

7.3 Choice- vs. Knowledge-based Assessment Across Levels

Similar as previous section, we first investigated whether each individual features we extracted from whole log files would predict students' final math test scores.

While on level 2, 13 out of 18 Catalog-related features are significantly predicted students' final math test scores, only 5 Catalog-related features are significantly predictors when using across levels. Among them, the most predictive feature is the same as using level 2: ResourceReviewDuration (the total duration that a Catalog is open). It significantly predicted students' final math tests scores: $\beta = 0.004$, $t(38) = 2.34$, p = 0.02. It alone also explained a significant proportion of variance in students' final math tests scores, R^2 = .13, $F(1, 38) = 5.46$, p= 0.02. So when using the whole game play logs, the best predictive Catalog-based feature is still the same as using the level 2 alone: ResourceReviewDuration. However, when considering the whole game play, the ResourceReviewDuration is less predictive than using the level 2 data alone.

Similarly, out of 13 Experiment-Room related features, the only feature that significantly predicted students' final math tests scores is again: GoalOrientedExperimentSuccessTry (the number of times students pick a color from the game board and then try to make the targeted color successfully in the Experiment Room). It significantly predicted students' final math tests scores: β = -0.45, $t(38) = -2.45$, p = 0.02. It alone explained a significant proportion of variance in students' final math tests scores: R^2 = .14, $F(1, 38) = 5.995$, p = .02. Again, when considering the whole game play, the GoalOrientedExperimentSuccessTry is less predictive than using the level 2 data alone: R^2 = 0.136, $p < 0.02$ vs. R^2 = 0.16, p = 0.01 respectively.

For the remaining three types of features, as when using the level 2 log alone, none of the features significantly predict students' final math test scores. In other words, again none of the embedded knowledge-based assessment on students' game play performance significantly predict students final math test scores. For example, PercCorrectGamePlay, the total number of times tile a student succeeded divided by the total number of tiles the student tried to play across the whole game, again did *not* significantly predict student's final math scores: R^2 = .02, $F(1, 38) = 0.88$, p = .35.

As with using level 2 alone, the choice-based assessment is a better predictor for students' math test school than the game-embedded knowledge-based assessment when using the single feature.

The brute-force search selects the best three features from all the four categories that would best predict students' final math test scores. Two features, WrongResourceReviewDuration (Catalog) and GoalOrientedExperimentSuccessTry (Experiment Room), are also shown in best three-feature model using Level 2 log alone; the other feature is:

> ExactDurationNoActivity (General): The total duration that a student is not involving any game playing activities such as reading Catalog, nor using Experiment Room, nor playing a game.

The results of the regression indicated the three predictors explained a significant proportion of variance in students' final math tests scores: R^2 = .34, $F(3, 36) = 6.27$, $p < .002$. All three features significantly predict students' final math tests scores (Table 2)

Table 2: Coefficient of Three Across Level Feature Model

	coeff	Sig
WrongResourceReviewDuration	0.008	$t(38) = 3.05$ p =0.004
GoalOrientedExperimentSuccessTry	-0.55	$t(38) = 3.20$ p =0.003
ExactDurationNoActivity	-0.005	$t(38) = 2.25$ p =0.03

In addition to the fact that both WrongResourceReviewDuration and GoalOrientedExperimentSuccessTry showed up in the best predicted models when considering level 2 alone and when considering the whole game play, in both models the former is positively correlated with students' school math performance while the latter is negative correlated. Additionally, note that the best model when considering level 2 alone beat the best model across level: R^2 =0.43 vs. R^2 =0.34.

When using the same three best features used in the level 2's best model to predict students' final math tests scores, the model is still significantly predict the student's school performance: R^2 = .25, $F(3, 36) = 4.01$, p = .01. We found that WrongResourceReviewDuration significantly predicted students' final math test scores (β = 0.006, p = 0.02), as did GoalOrientedExperimentSuccessTry (β = -0.44, p = 0.02), but not TryErrorPick, (p = 0.95)

Again, note that none of the game-play performance related features were included in the final best three-feature model. This again suggests that when considering the whole game play, the choice-based assessment still be a better predictor than the game-embedded knowledge-based assessments.

Furthermore, our results also suggested that the choice-based assessment when using level 2 alone is more predictive than the choice-based assessment using the whole game play. So it suggested that it is important to note that for certain skills, the choice-based assessment should be nimble. 5 minutes on Lightlet is more efficient to detect effective learners than asking student to spend 20 minutes on it.

Finally, we have shown that choice-based assessment is more predictive than game-embedded knowledge-based assessment. One more interesting question to answer is: did the choice student made during the game play help their performance in the game? There is insufficient space in this paper to go into detail. But our overall finding is that the choice student made during the game play indeed significantly predicts their performance in the game.

For this analysis, we treat the third level of Lightlet as a posttest for level 2. If students make good learning choices on level 2 and learn about additive color, then they should do well on level 3. It turns out that the same choice pattern in level 2 that predicted learning in students' math class also predicted performance on level 3 (how efficient they clear the game board on level 3): R^2 = .32, $F(3, 36) = 5.61$, $p = .003$. We found that only one feature WrongResourceReviewDuration significantly predicted student's level 3's performance ($\beta = -0.19$, $p = 0.0004$) while the other two were not significantly predictive: TryErrorPick, ($p = 0.10$) GoalOrientedExperimentSuccessTry ($p = 0.15$).

To summarize, students who chose to explore the Catalog were more likely to do better in school. In fact, the amount of time students committed to figuring out the Catalog entries shortly after the Catalog became available predicted about 43% of the variation in the students' grades in their mathematics classes (the only classes for which we had records). In other words, all students tried to level-up in our game, but those who chose not to engage in critical thinking while doing so were also doing worse in school.

8. Conclusions

For many, assessments are a lighthouse in the fog of education—a clear guide by which to make safe decisions. But in reality, assessments create the fog. Current assessments perpetuate beliefs that the proper outcomes of learning are static facts and routine skills—stuff that is easy to score as right or wrong. Interest, curiosity, identification, self-efficacy, belonging, and all the other goals of informal learning cannot even sit at the assessment table, because these outcomes are too far removed from current beliefs about what is really important.

Assessments seem to be built on the presupposition that people will never need to learn anything new after the test, because current assessments miss so many aspects of what it means to be prepared for future learning. These frozen-moment assessments have influenced what people think counts as useful learning, which then shows up in curricula, standards, instructional technologies, and people's pursuits.

Teachers may tell students about the importance of persistence, critical thinking, interest development, and a host of other keys to a successful life. But tests provide the empirical evidence that students use to decide what is truly valued. If an assessment focuses on the retrieval and procedural application of narrow skills and facts, this is what students will think counts as useful learning. By changing assessments to concentrate on choices, we should be able to improve beliefs about what constitutes useful learning.

If the fog were lifted, we would see that most of the stakeholders in education care first and foremost about people's abilities to make good choices. Making good choices depends on what people know, but it also depends on much more, including interest, persistence, and a host of twenty-first-century soft skills that are critical to learning. Where we can anticipate a stable future—decoding letters into words is likely to be a stable demand for the next fifty years—then knowledge- and skill-based assessments make sense. In relation to those aspects of the future that are less stable, though, people will need to choose whether, what, when, and how to learn. Hence, it is important to focus on choices that influence learning, and assessments should measure those choices. Choice is the critical outcome of learning, not knowledge. Knowledge is an enabler; choice is the outcome.

Assessing choices during learning has a number of attractive properties. Foremost, choice-based assessments are process oriented. They examine learning choices in action rather than only the end products. This process focus makes it possible to connect the learning behaviors during the assessment to processes that occur in a learning environment. Second, the assessments reveal what students are prepared to learn, so they are prospective as opposed to retrospective. Third, choice resonates with the rest of the social sciences that examine the movements of people, money, and ideas. Fourth, choices do not lend themselves to simplistic reifications whereby things like people's knowledge or personality traits are misinterpreted as independent of context and immune to change. Fifth, choices can measure a much greater range of learning outcomes than fact retrieval and procedural application. Sixth, learning choices are a good candidate for inclusion in standards, which currently define what knowledge students should have but stay strangely silent about the processes of learning themselves.

Recent advancements in technology create a special opportunity for moving toward a new paradigm of assessment. There are risks, however. People may only use technology to make us faster and more entrenched in doing the wrong thing. When used well, technology makes it possible to create and validate choice-based assessments by using the rapid generation of interactive environments, crowdsourcing, automated logging, and educational data mining. Thus, it is possible for choice to become the core of assessment (and not in the degraded sense of multiple-choice tests). In this paper, we provided an anchoring example of a computerized, choice-based assessment, Lightlet.

Tracking the process of learning is different from simply detecting whether a student knows an answer or not, which is the output of most tests. Choice-based assessments can provide a much richer corpus of information from which to draw actionable information about learners. We can locate the source of the problem rather than just the consequence. The students who were doing the worst in math class, for instance, were those who used the Experiment Room to solve each problem through trial and error. These students, instead of trying to develop an overall understanding of additive color, were simply attempting to get the right answer for each problem in turn. In the best case, identifying this pattern of choices can help a teacher address the underlying learning issue, which is that the students are trying to solve each problem in turn rather than discovering the general principle that governs the solutions to all problems.

In an initial study using a similar environment with sixth grade children, the results were quite clear. Children who chose to look at the Catalog of charts during the game were doing better in school. In fact, the different choice patterns students committed during the game play predicted about 43 percent of the variation in the students' grades in their mathematics classes. While all the students seemed happy to play the game, those who chose not to engage in critical thinking were also the students who performed worse in mathematics. The 43 percent level of prediction is high, especially considering that Lightlet has little to do with solving math problems as they appear on the children's mathematics tests. The assessment captured something crucial about how these children go about learning that is affecting their success in mathematics—and will likely do so in the future.

Overall our results offer two take-home messages. First, by assessing students' choices during game playing, we can discover a great deal about the processes they do or do not use to learn. Second, we can assess choices that are critical to learning, but that are missed by most tests.

To summarize, with more choices and interactivity comes more information about the learner. Performance assessments, such as portfolio and project-based assessments, have tried to capitalize on the increased information found in choice-rich environments (e.g., Resnick and Resnick 1994). Richard Shavelson, Gail Baxter, and Jerome Pine (1991), for example, describe a kit-based performance assessment for science. Students conduct physical experiments to determine which brand of paper towel absorbs more water. The assessment provides information about the students' abilities (or inclinations) to use experimental logic and take careful measurements. Unfortunately, the authors also point out that performance assessments can be prohibitively expensive to deploy and score at scale. Technology can help overcome the difficulties associated with increased information. Computers can deliver assessments where students make choices about how to learn, and the computers can automatically log all user behaviors that might be of interest to a teacher, assessor, or researcher, ranging from chat logs to virtual interpersonal distance to direction of gaze. It is an ethnographer's thick description for free. Computers provide new efficiencies that make tractable what was once impracticable. And with new empirical capabilities, new theories are sure to follow.

People generally in system performance measure for predictive analysis. Our research, however, shows that behaviors features can be far more informative

9. REFERENCES

[1] Aleven V., Koedinger, K. R., & Popescu, O. (2003). A Tutorial Dialog System to Support Self-Explanation: Evaluation and Open Questions. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), Proceedings of the 11th International Conference on Artificial Intelligence in Education, AI-ED 2003 (pp. 39-46). Amsterdam, The Netherlands: IOS Press. Finalist for Conference Best Paper Award, AIED 2003

[2] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540.

[3] Bransford, J. D., and D. L. Schwartz. 1999. "Rethinking Transfer: A Simple Proposal with Multiple Implications." Review of Research in Education 24:61–100.

[4] Chi, M., Koedinger, K. R., Gordon, G., Jordan, P. W., & VanLehn, K. (2011). Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. C. Stamper (Eds.), Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, July 6-8, 2011 (pp. 61–70).

[5] Chi, M., VanLehn, K., Litman, D. J., & Jordan, P. W. (2011a). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. User Model. User-Adapt. Interact., 21(1-2), 137–180.

[6] Doris B. Chin, Ilsa M. Dohmen, Britte H. Cheng, Marily A. Oppezzo, Catherine C. Chase, and Daniel L. Schwartz. (2010) Preparing students for future learning with Teachable Agents. Educational Technology Research & Development.

[7] Feuerstein, R. 1979. The Dynamic Assessment of Retarded Performers: The Learning Potential Assessment Device, Theory, Instruments, and Techniques. Baltimore, MD: University Park Press.

[8] Gee, J. P. 2003. What Video Games Have to Teach Us about Learning and Literacy. New York: Palgrave.

[9] Hogyeong Jeong, Gautam Biswas (2008) Mining Student Behavior Models in Learning-by-Teaching Environments, 127-136. In The 1st International Conference on Educational Data Mining.

[10] Iyengar, S. S., and M. R. Lepper. 1999. "Rethinking the Value of Choice: A Cultural Perspective on Intrinsic Motivation." Journal of Personality and Social Psychology 76:349–366.

[11] Schwartz, D. L., and J. D. Bransford. 1998. "A Time for Telling." Cognition and Instruction 16:475–522.

[12] Schwartz, D. L., and T. Martin. 2004. "Inventing to Prepare for Learning: The Hidden Efficiency of Original Student Production in Statistics Instruction." Cognition and Instruction 22:129–184.

[13] Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. Daniel L. Schwartz, Catherine C. Chase, Marily A. Oppezzo, & Doris B. Chin. I (2011). *Journal of Education Psychology.*

[14] Stevens, R. H., & Thadani, V. (2007). Quantifying students' scientific problem solving efficiency and effectiveness. Technology, Instruction, Cognition and Learning, 5(4), 325-337.

[15] Vygotsky, L. S. (1934) 1987. The Collected Works of L. S. Vygotsky, ed. R. Rieber and A. Carton. New York: Plenum.