Got Game? A Choice-Based Learning Assessment of Data Literacy and Visualization Skills

Doris B. Chin, Kristen P. Blair, and Daniel L. Schwartz

Stanford Graduate School of Education

Phone: 650-725-5480

Fax: 650-725-5916

Email: dbchin@stanford.edu

URL: aaalab.stanford.edu

Acknowledgements:

This material is based upon work supported by the National Science Foundation under Grant numbers 0904324 and 1228831, the John D. and Catherine T. MacArthur Foundation, and the Gordon and Betty Moore Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies. The authors would like to thank Jacob Haigh and Neil Levine for their key contributions in the development of the assessment game, as well as Rochelle Urban and Megan Schufreider for their work on the pilot curriculum and study.

ABSTRACT:

In partnership with both formal and informal learning institutions, researchers have been building a suite of online games, called choicelets, to serve as interactive assessments of learning skills, e.g. critical thinking or seeking feedback. Unlike more traditional assessments, which take a retrospective, knowledge-based view of learning, choicelets take a prospective, process-based view and focus on students' choices as they attempt to solve a challenge. The multi-level challenges are designed to allow for players' "free choice" as they explore and learn how to solve the challenge. The system provides them with various learning resources, and tracks whether, what, how, and when they choose to learn. This paper briefly describes a partner's curriculum focused on data literacy and visualization, the design of a choice-based assessment for their program, and reports on an initial study of the curriculum and game with 10th grade biology students. Results are presented in the context of the design research questions: *Do student choices in the game predict their learning from the game? Does the curriculum teach the students to choose more effectively with respect to data literacy?* Future work for choice-based assessments is also discussed.

KEY WORDS:

Educational technology; learning assessment; educational assessment; game-based assessment; science education

1 Introduction

How should we measure learning? There is a growing call to shift the focus of assessment from a retrospective, mastery model to a more prospective, process-based model (Schwartz & Arena, 2009; Shute *et al*., 2009). This call hails in large part from the growth of technology as an increasingly ubiquitous and driving force for the discovery of new knowledge. Knowledge must be treated as a fluid entity, rapidly and constantly evolving, and our children must learn how to navigate this river of information as independent travelers, without the guidance of their teachers or parents. Therefore, the argument is that we should assess not only *what have kids learned*, but also *what are they prepared to learn* as they encounter new challenges in the future.

The move toward a more prospective, process-based view of learning is reflected in the newly adopted Common Core State Standards and Next Generation Science Standards, both of which explicitly incorporate standards of practices and skills, e.g. "integrating multiple sources of information" and "obtaining, evaluating, and communicating information" (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; NGSS Lead States, 2013). Learning skills and practices are also the focus of such organizations as the Partnership for 21st Century Learning (http://www.p21.org/), as well as in reports by the National Science Foundation (Friedman, 2008) and the National Research Council (2009).

The current landscape of computer-based assessment protocols, however, lags behind new process-oriented standards. Most assessments implicitly embrace a mastery framework for educational outcomes, and typically sample from a subset of facts and procedures to infer students' command of the full set of desired knowledge. Even newer assessments, such as Smarter Balanced, follow a similar model. They include better probes of student understanding, but they remain retrospective measures of what students have mastered from a bounded curriculum.

This mismatch between current assessments and the goal of preparing students to continue learning is exceptionally problematic for evaluating informal learning experiences. Informal experiences are, by design, relatively unbounded to allow for the free choice of learners. They are rarely designed on the assumption that students will master the same content in the exact same way. Moreover, the aim of such experiences is to put individuals on a trajectory of continued, life-long learning; relevant learning goals go beyond simply knowledge accumulation, *per se*, but also include developing skills, fostering engagement, as well as changing behaviors and attitudes of learners (cf. Falk & Dierking, 2002). The design for idiosyncratic experience and non-factual outcomes has been an assessment challenge. Evaluations of informal learning experiences frequently rely heavily on self-report surveys and observation protocols. These methods can provide rich data, but they are time-consuming and have difficulty tracking effects beyond the immediate administration of the informal experience (Friedman, et al., 2008; NRC, 2009).

Game-based technologies are one possible solution to bridge both formal and informal learning environments and assess learning processes that are active across settings. There has been much work and discussion on how the affordances of computer and video games allow for interactive, "just in time" learning (cf. Gee, 2003; Garris, Ahlers & Driskell, 2002; Van Eck, 2006). Games, it is argued, are not only ideal vehicles for students to pursue their own trajectories of learning, but they also provide the means to evaluate students in a context of whether, what, how, and when they choose to learn (Mayo, 2009; Schwartz & Arena, 2009; Shute, 2011; Ifenthaler, Eseryel, & Ge, 2012; Schwartz & Arena, 2013; Shute, et al., 2015).

2 Choice-Based Assessments

In partnership with several informal and formal learning institutions, researchers have been building a suite of online games to serve as interactive assessments of learning processes and skills (Schwartz & Arena, 2013; Chi, Schwartz, Chin, & Blair, 2014; Conlin, Chin, Blair, Cutumisu, & Schwartz, 2015; Cutumisu, Blair, Schwartz, & Chin, 2015). These games, called choice-based assessments ("choicelets" for short), capture and highlight the importance of the decisions that children make as they learn. Each choicelet targets a learning process of interest, e.g. critical thinking or trying multiple solutions. Choice-based assessments follow three broad design principles:

1. Typical Performance vs. Maximal Performance

Tests and high-stakes assessments frequently trigger "maximal performance" behaviors in students which do not accurately reflect their everyday learning behaviors. (Sackett, Zedeck, & Fogli, 1988; Klehe & Anderson, 2007). Moreover, such assessments may match learner expectations in school contexts, but they are decidedly not fun, and thus undesirable in informal learning contexts. Therefore, researchers design assessments that are games which explicitly do *not* feel like tests, under the assumption that typical performance is a better proxy for children's independent learning behaviors.

2. Preparation for Future Learning

Each game presents players with a multi-level challenge that they must solve. Embedded within the game are learning resources that players may choose to explore (or not) as they attempt to solve the challenge. Thus, by design, there is something new to learn within the game, and successfully completing the game does not rely solely on prior content knowledge that players bring into the game.

Choice-based assessments are not intended as embedded assessments for a particular curriculum. Rather, each game is designed as a stand-alone assessment, compatible with a range of curricula or learning programs. This is in contrast to other game-based assessments, for example, Shute, Ventura, Bauer and Zapata-Rivera (2009), who employed a stealth assessment that effectively tracked student persistence within a computer environment called, Newton's Playground. The assessment was embedded within the game, which required multiple hours of play. Embedded assessments are a powerful way to track student performance as they move through a curriculum, and the authors showed that game could automatically adapt based on student performance. Embedded assessments, however, are not ideal for comparing the effectiveness of different curricula or treatments, because they are integrated within a specific curriculum and they take a long time to complete. Choice-based assessments are meant to evaluate whether a curriculum prepares students to learn new content, not originally covered in the curriculum, and as such, they do not need to be tied to a specific curriculum. They are also designed to take about ten to fifteen minutes.

3. Free Choice

For each game, there are multiple and, at times, competing avenues by which players can reach their goal. Each built-in learning resource can help the players in their quest, but the challenges are designed such that players can advance in the game by choosing different resources, in different order, at different times, or even not using the resources at all. (Trial-and-error is not an uncommon strategy employed by players.) This differentiates choice-based assessments from most computerized instruction and games, where there is only one path to success. The free-choice principle makes it possible to determine how and whether players choose to learn when they are not being forced or coerced by the structure of the environment.

Choice-based assessments entail the explicit assumption that some choices are better than others for learning in a given context. This creates a special burden for the design of choice-based assessments. Unlike the claim that 2+2=4 is better than 2+2=5, there is no *a priori* claim that one choice is better than another or that a choice can be taught. One needs to prove it empirically, often for each instance, because researchers have tended to treat student choice as a self-selection confound to be eliminated rather than the construct of interest. Therefore, the first empirical step in the design of a choice-based assessment is to determine whether specific choices correlate with better learning and whether they can be influenced through instruction. The current study reflects this first step, which often depends on exploratory data analyses to determine the best sources of signal in the game. The second step, which the current research has not taken yet, involves replicating the original finding to ensure it is stable and not spurious.

To demonstrate the learning value of a given choice, researchers can use in-game measures and external learning measures, such as a posttest on the content within the game. As an example, a choicelet called Posterlet provides players with choices of either positive or negative (constructive) feedback on posters they have designed in the game. Both types of feedback contain *equivalent* information about graphic design principles. While prior studies had found that constructive criticism is better for learning, the prior studies had not matched the informational equivalence of positive and negative feedback messages (positive feedback messages tended to be non-specific praise). Critically, they had never examined what happened when people actually had the opportunity to choose their own feedback. Therefore, the researchers first used Posterlet to determine whether seeking constructive feedback is indeed better for learning. In studies with players ranging from middle-school to college students, they found that seeking more constructive criticism correlated with improved posters in the game. Complementing the in-game measure of learning, the researchers also found that seeking more constructive criticism led to better performance on an external posttest of graphic design principles (Cutumisu et al., 2015).

3 Storylet: An Example of a Choice-Based Assessment

This paper presents an instantiation of a choicelet that was developed in partnership with an informal learning institution and piloted with 10th grade biology students. The first section describes the partner's program and their desired learning outcomes. The second section outlines the design process for a choicelet compatible with the targeted learning outcomes. Next, a pilot study and initial results are presented. Design research questions that are addressed with the study include:

1. *Do student choices in the game predict their learning from the game?*
2. *Does the curriculum teach the students to choose more effectively with respect to data literacy and visualization skills?*

Finally, implications and future directions for the work are discussed.

3.1 Instructional Background

To make good decisions or choices, individuals need to know how to sift through an ever growing mountain of information. Data literacy has, arguably, become an important life skill, and components of this skill can be found scattered throughout the Common Core State Standards for both Math and English Language Arts, as well as the Next Generation Science Standards disciplinary practices (Table 1).

[Insert Table 1 here]

In support of these new standards, staff at an internationally recognized science center in California developed a week-long curriculum focused on data literacy. The intended audience was high school science and math classes and out-of-school programs. "Storytelling with data" was the premise of the instruction; teaching students how scientists communicate with data and instilling in them good graphic principles for visualization.

Through a combination of instructional methods, including lecture/discussion, worksheets, and collaborative work, the curriculum focused on two main principles for data visualization which were presented to students:

1. Simplicity
    a. Visualizations should aim to impart one or two key messages.
    b. Everything on a graphic should have a reason for being there.
2. Truthfulness
    a. Data must be accurately represented and labeled.

Key activities involved comparing and contrasting data visualizations in a variety of domains (see Figure 1 example). Students were asked to consider what central idea the various graphics were trying to convey, as well as the methods used by their authors to communicate data.

[Insert Figure 1 here]

3.2 Design of Storylet

Researchers collaborated on the project to help assess learning outcomes for the curriculum. The timeframe was such that the choicelet was designed in parallel with the curriculum, and researchers met multiple times with science center

staff over a period of several months. It was agreed that the assessment should not focus on the specific content presented in the program *per se*, but on whether the curriculum prepared students to better "tell stories with data."

A major challenge for the design of game or simulation-based assessments is the translation of a curriculum's learning goals into a *measurable* construct or learning strategy of interest. Mislevy and colleagues have formulated a framework for the design of assessments called Evidence Centered Design (Mislevy, Almond, & Lukas, 2003; Mislevy, 2011). The three main components or models of the ECD framework include the competency model, which asks the question, "What knowledge or learning skill should be targeted?" The evidence model asks, "What observable evidence, e.g. actions or products, would serve as indications of competency for the targeted construct?" The task model asks, "What situations would elicit these desired actions or products?" Though researchers did not explicitly follow ECD, it may be useful to map some of the current, choice-based design decisions onto familiar ECD landmarks in the following description.

The science center staff's focus on "telling stories with data" was translated by researchers as most closely resembling the NGSS practice 8 of "obtaining, evaluating, and communicating information" (see Table 1). More specifically, this skill (or competency model, to use ECD terminology) includes the abilities to 1) determine the central idea from scientific text, and 2) compare, integrate and evaluate sources of information presented in visual form (e.g. graphs or tables). The curriculum designers delved yet further by framing the evaluation of data visualizations along the two principles mentioned in section 3.1 – simplicity and truthfulness.

Over the course of approximately 10 weeks, researchers developed a simple choicelet game, dubbed "Storylet." The choicelet is based on the rather fanciful and hopefully fun premise that players have been hired as junior editors at the OhNo Gazette. Their challenge is to peruse and choose appropriate graphics and titles for three different stories. Thus the task model ensures that behaviors representative of the chosen learning strategies are present and available to players. For each story, there are three main activity tabs (pick photo, pick chart, and read story), as well as two minor activities (pick title and review). Figure 2 illustrates the flow of possible choices between tabs in the game, beginning with "Choose Topic" in the upper left.

[Insert Figure 2 here]

As framed by the PFL and free-choice design principles, each of the three main activity tabs (photo, chart, and story) represents a learning resource, and the game allows players to freely explore amongst them. For each story, students have six possible choices for photo and four possible choices for chart (as well as five possible titles). They are presented in thumbnail form, and players can choose to zoom in for a closer look on each graphic. After choosing a chart and photo, players are allowed to see the titles tab, where they must select an appropriate headline. Their final step for each game level is to click "Go to press" to review the published story (along with a humorous advertisement at the bottom and reader comments). Up to this point, players are still allowed free access to all activity tabs. Critically, players can make their editorial selections without ever going to the story tab. Once learners submit their final story, they see a "buzz meter" that provides students with feedback on the appropriateness of their editorial decisions. The buzz meter was intended as a straightforward measure of game performance and learning, but in the ECD framework, the underlying scoring rubric can be viewed as part of the evidence model that distinguishes different

levels of competency: the charts, photos, and titles are all scored for how well they capture the central idea of the story; additionally, the charts are graded as good or bad representations of the data visualization principles, e.g. some charts are clearly labeled, others are not. At the bottom of the review page is a "Select your next story" button that players may click when they are ready to move on.

Log files from the system provide a continuous clickstream with time stamps for students' choices as they explore the tabs and step through the three topics. The hypothesis was that the data visualization curriculum would tilt those students toward different patterns of behavior and learning compared to their control peers, and these differences would manifest most strongly in their graphic-related choices, e.g. spend more time on the graphics tabs, zoom in on more charts.

Given the constraints of the development timeline, researchers did not build affordances into this first version of the game for learners to explicitly critique the graphics or create their own data visualizations. Rather the focus was on finding each story's central idea, then evaluating the available images and data representations to choose the ones most suited for the story. As an example, the text for one of the three stories, *Argentine Ant Attacks*, is shown below. Figure 3 displays the chart choices for this story.

> *Two species of ant are waging war in California. In the late 1800's, the Argentine ant stowed away on cargo ships from South America and invaded the world. They are only 2mm in size, but they make huge colonies wherever they go, wiping out native ants. However, a California ant is fighting back! Stanford scientists recently discovered that the local winter ant has a secret weapon. It releases poison from its tail! The winter ants only use the poison in self-defense. They mostly release the chemical when other ants are within 30cm of their nest. Scientists also pitted winter and Argentine ants in small gladiator-like fights. Sometimes, the percentage of winter ants was 80% compared to 20% Argentine ants. Sometimes it was 50-50, or 20-80. The winter ants were over 15 times more likely to use their poison when badly outnumbered. The poison kills enemy ants about 79% of the time, helping California winter ants win back their homeland.*

[Insert Figure 3 here]

4 Research Study & Methods

Participants were 10th grade biology students from a California public high school (N=93; Table 2; researchers did not have access to class grades or standardized achievement scores). One of the school's teachers volunteered to participate in the pilot with students from three of her classes (N=49). These "Data Viz" students underwent the data visualization lessons (six instructional hours, carried out by science center staff over roughly three weeks). Control students were from her remaining class, as well as two classes from the school's other biology teacher (N=44). Logistical constraints required that all classes remain intact.

[Insert Table 2 here]

Two weeks after the last day of instruction, all students played the Storylet game for ~10 minutes during their regular biology class. At the end of the game, students took a discrete assessment of what they learned from the game. There were five questions that targeted factual story content and design principles demonstrated in the game graphics (Table 3). Both types of questions were included to parse out the graphics learning and to determine if there were selective treatment effects. It was expected that both groups of students would do equally well on the factual questions, but that the Data Viz students would do better on the graphics-focused questions. The post-test questions were included to provide learning measures external to the game. In combination with game performance scores (buzz meter), the different sets of learning measures can be analyzed to evaluate whether certain choices (e.g. the decision to look frequently at the data graphics) do, in fact, lead to better learning.

[Insert Table 3 here]

5 Results

The pilot study was used to evaluate both Storylet and the curriculum. The first set of analyses addresses whether some student choices are better for learning than others. These exploratory analyses are conducted on data from students who played Storylet to completion, regardless of condition. The second set of analyses focuses on the impact of the curriculum - does undergoing the data visualization curriculum have a measurable effect on student choices and learning in the game?

5.1 *Do student choices in the game predict their learning from the game?*

To establish that the choices players make in the assessment have an influence on how well they learn, data was extracted for the three main game activities: choose chart, choose photo, and read story. Metrics extracted include the number of times an activity tab was visited, time spent in each tab, and number of times specific charts or photos were "picked" for closer examination. The extracted data were used in stepwise regressions to predict students' post-test scores, without regard to their treatment condition.

The portion of the posttest that specifically tested the factual scientific content was analyzed first (Table 3, questions 1-3). To control for overall time on task, total game time was entered into the regression ($F_{(1,85)}$= 4.17, p=.044, R^2=.047). The next step of the analysis determined that, even controlling for total game time, time spent reading was a significant predictor of science content learning: F-change$_{(1,84)}$=5.2, p=.025, R^2-change=.06; final model, $F_{(2,84)}$=4.79, p=.011, R^2=.10. Thus, as expected, the choice pattern of spending *proportionately more time* on the reading was important for learning factual content.

Of more interest was the question of what choice data would predict performance on the questions that tested for understanding of the graphical principles emphasized in the game (Table 3, questions 4-5). As with above, total

time spent in the game was entered into the regression, although this time it was not significant ($F_{(1,85)}$= .33, p=.57, R^2=.004). It was expected that a chart-specific metric would next enter the regression, because the charts held the information relevant to the graphic principles. Surprisingly, the total number of "graphic" picks (times when player clicked for closer looks on charts and photos) was the only other choice predictor to enter the equation: F-change$_{(1,84)}$=6.8, p=.011, R^2-change=.075; final model, $F_{(2,84)}$=3.57, p=.03, R^2=.08. Thus, choosing to visit the graphics correlated with better understanding of graphical principles. The lack of a chart-specific correlation is considered in the discussion.

To determine whether game choices also correlated to performance measures within the game, researchers analyzed the buzz meter, which tracks the quality of a player's final publication for each story. Each story had a maximum buzz meter score of 4 points: 2 possible points for a correct chart (1 point for main message + 1 point for good graphic principles), 1 point for a correct photo, and 1 point for a correct title. This score represents in-game learning outcomes of specific interest to the curriculum designers: students had to read and understand the central idea of the story, as well as compare and evaluate the graphical and title options to choose the best-fitting options.

The researchers conducted a stepwise regression on students' total buzz meter score (12 points maximum: 3 stories x 4 points), using all available log data as possible predictors. We found that four game metrics entered into the equation, each linked to a different activity in the game. In order, the variables that entered were: total story time, $F_{(1,91)}$=11.2, p=.001, R^2=.11; total photo time, F-change$_{(1,90)}$=6.2, p<.02, R^2-change=.06; average chart tab time, F-change$_{(1,89)}$= 4.3, p < .05, R^2-change=.04; average title pick time, F-change$_{(1,88)}$=6.6, p<.02, R^2-change=.06; final model, $F_{(4,88)}$=7.8, p<.001, R^2=.26. The analyses indicated that the entered variables all showed positive correlations with the buzz meter score, *except* average title pick time, which was negatively correlated. One possible interpretation of this is that students who better understood the main message of each story could more quickly pick out the best title, while their peers took longer to mull their choices.

An examination of the correlations between the post-test and buzz meter scores (Table 4) reveals strong positive correlations between the graphics-focused post-test items and all components of the buzz score. There are no significant correlations with the content-focused post-test items, except with the title score. Thus, the buzz meter of Storylet, which is visually available to the students after each publication, appears to be specifically targeting the data literacy skills of evaluating and communicating the main point.

[Insert Table 4 here]

Lastly, the researchers addressed a main goal of their partner's Data Viz curriculum. The curriculum was designed to teach students to look for the main message of a story. *Post-hoc*, researchers wanted to know if they could predict, *using only players' in-game choices*, if students were reading for "factual" content or "overall message." To parse this out, multivariate analyses were conducted on the post-test content measures and the buzz meter score. The post-test content measures (Table 3, items 1-3) specifically tapped factual knowledge, while the buzz meter score tracked if students understood the main message of the stories. Data analyses indicate that total story time predicted both factual learning and message learning ($F_{(1,84)}$=11.1, p=.001 and $F_{(1,84)}$=8.4, p=.005, respectively). However, total

photo time was selectively predictive only for message learning ($F_{(1,84)}$=4.7, p=.03), and not for factual learning ($F_{(1,84)}$=1.9, p=.17). The data suggest that by looking at students' non-reading related choices, the game can help differentiate whether players are reading to find the main message. Those students who spent proportionately more time perusing and choosing the best photos were also more likely to have understood the main message. This hints, though does not prove, that these students were indeed reading for the overall message versus for factual details. If true, this may offer a simple way to track reading strategies in on-line environments, namely, measure whether people are looking at supporting visualizations and not just the text itself.

In summary, data from Storylet indicate that students' choices *do* correlate with their learning outcomes, including content and graphics-focused questions on a post-test, as well as in-game learning measures. It must be clarified, however, that the first two results are based on post-test measures only. Learning gains could not examined because instructional time constraints did not allow for administration of a pre-test. It is possible that individual differences in the measures of learning may be the result of incoming knowledge rather than in-game learning. In addition, though system metrics were able to differentiate learning outcomes, the results indicating the relative importance of photo-related choices were somewhat surprising. The Discussion addresses possible reasons for this finding.

5.2 *Does the curriculum teach the students to choose more effectively?*

These next set of analyses focus on the impact of the curriculum – did the data visualization lessons lead students to make better (or at least different) learning choices in Storylet than their control peers? As a preliminary check, total time spent in the game was analyzed. There was no significant difference between the control and data visualization students ($t_{(91)}$=.39, p=.70), thus any differences between groups are likely not due to overall time on the Storylet task.

Next, the analyses focus on students' tab activity, specifically the number of times they visited the main tabs for examining graphics and for reading. (In the preceding analyses, the stepwise analyses chose variables to maximize explained variance. This led to a mix of total time, average time, and frequency of tab choice. However, the total and average amount of time a player spent on a particular tab was highly correlated with how frequently the player clicked on the tab to see the relevant content. Therefore, the current analyses use tab frequency to maintain conceptual consistency with negligible loss in variance explained.) A 2x2 repeated-measures analysis crossed the between-subjects factor of treatment (Control vs. "Data Viz") by the within-subjects factor of tab type (Graphics vs. Stories). Graphic picks (chart and photo) and story reading were significant in the preceding analyses. Figure 4a shows the mean number of tab visits by condition.

[Insert Figure 4 here]

The analysis found a large main effect for tab type, as expected ($F_{(1,91)}$=356.75, p<.0001, partial η^2 =.797). There was one passage and many graphics for each story, so it makes sense that there were more clicks on the graphics tabs. There was a marginal main effect for treatment group ($F_{(1,91)}$=3.58, p=.062, partial η^2 =.038). Of more interest, there was a significant interaction of treatment x tab activity ($F_{(1,91)}$=5.19, p=.025, partial η^2 =.054). Students from both

conditions chose to read the stories about the same number of times, but the Data Viz students visited the graphics tabs (charts & photos) more often than their control peers. The curriculum designers and researchers predicted that the Data Viz students would come back to the story tabs more often, so they could better understand the main points to help pick the best graphics. However, this was not the case.

A similar 2x2 repeated-measures analysis was conducted on the post-test measures, crossing the between-subjects factor of treatment (Control vs. Data Viz) by the within-subjects factor of post-test question type (Graphics vs. Factual content). The preceding analyses independent of condition found that students who perused the graphics more (clicked more often on individual graphic images) did better on the graphics posttest questions, and that students who spent proportionally more time on the stories did better on the content learning. Combined with the analysis showing that the Data Viz students clicked on graphics more frequently, one would predict that Data Viz students would do better on the graphics portion of the posttest, and that control students would do better on the content portion of the test. This is, in fact, what happened, as evidenced by Figures 4b. There was a significant interaction effect, albeit of smaller effect size ($F_{(1,85)}$=4.4, p=.04, partial η^2=.049). The analysis also found a significant main effect for question type ($F_{(1,85)}$=18.2, p<.001, partial η^2=.177), indicating that students did better on the content questions compared to the graphics questions. There was no main effect for condition ($F_{(1,85)}$=.04, p=.85).

In summary, the pilot study provides evidence that Storylet detected effects of earlier instruction: exposure to the data visualization curriculum affected how the treatment students played the game – as measured by proportion of time and click patterns on different game elements – as well as what they learned – as measured by the buzz meter and post-test scores. These claims must be viewed with some caution, however, based on the logistical necessity of recruiting a volunteer treatment teacher with intact classes. Teacher effects could be a possible confound. Moreover, the findings are based on mining the data, rather than *a priori* specification of the relevant variables. Storylet was designed to detect the effect of instruction on student choices to read and look at graphic material, but precisely where the signal would show up most strongly required *post-hoc* analyses. Therefore, the work requires replication now that the relevant variable combinations have been found. This would also address concerns with multiple comparisons.

6 Discussion

The impetus behind many of the calls for 21st century skills is to ensure that students are prepared to learn on their own — that they can both solve novel problems and learn new content areas without the guidance of their teachers or parents (Schwartz, Bransford, & Sear, 2005). Assessments should be aligned to this goal. Currently, researchers are working to develop a suite of choice-based assessments that captures children's learning behaviors, including games that target intelligent persistence (Schwartz and Arena, 2013), critical thinking (Chi et al., 2014), design thinking skills (Conlin et al., 2015), choosing to seek feedback and to revise (Cutumisu et al., 2015), and choosing to organize information (Semmens, Blair, & Schwartz, in prep). They have found that students' in-game choices do, indeed, predict differences in their learning outcomes. In addition, researchers have found evidence for the diagnostic value of these types of assessments: students' choices within the game are impacted by their learning experiences and correlated with external measures of learning.

An important direction for future work with this choicelet and others is to ask the question: *Is the information useful?* Does the information allow researchers and educators to (a) compare different models of instruction, (b) provide feedback which can help them improve their curricula, and (c) identify children who could benefit from instructional intervention. For the data literacy and visualization project, the results indicate that the curriculum successfully helped students to better understand graphical principles, and more importantly, led them to better extract the main message of the stories (albeit somewhat at the sacrifice of factual learning). Some of the results indicate that the learning effects may have not been isolated to evaluating good data visualizations *per se*, but also heightened student focus on *photo* images as well. In particular, the analyses for the graphics post-test questions indicate that the most significant predictor for performance was the *combined* number of charts and photos picks, not solely charts, and the analyses for game performance (buzz meter scores) showed a photo-related metric entering the regression before a chart metric. These findings could stem from the curriculum or its implementation; many of the lessons used colorful and visually appealing infographics that, by design, straddled the line between pictures and data representations. Additionally, framing the lessons as "telling stories" with data may have naturally triggered an increased awareness of *all* types of images including photos. An alternative, though not mutually exclusive, explanation for the results could be in the design of Storylet, perhaps in the presentation of the task or its difficulty - the stories may have been too long or the charts may have been difficult to interpret. Researchers hope to further investigate these design issues by repeating the classroom experiment and conducting more in-depth interviews with students as they play. The next iteration of the game will aim to build in more interactions around the charts to allow better "triangulation" on the targeted data literacy skills, e.g. a graphing module in which students could translate data from tables into visual charts or infographics. Another modification will attempt to increase the sophistication of Storylet by allowing teacher-driven content to be integrated into the game, more closely aligning the graphical choices with the design principles and data literacy.

References

Chi, M., Schwartz, D. L., Chin, D. B., & Blair, K. P. (2014). Choice-based assessment: Can choices made in digital games predict 6th-grade students' math test scores? In *Proceedings of the 7th International Conference on Educational Data Mining*, 36-43.

Conlin, L.D., Chin, D.B., Blair, K.P., Cutumisu, M., & Schwartz, D.L. (2015). Guardian angels of our better nature: Finding evidence of the benefits of design thinking. *Proceedings of the American Society for Engineering Education, June 2015*, Seattle, WA.

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2015). Posterlet: A game-based assessment of children's choices to seek feedback and to revise. *Journal of Learning Analytics, 2(1):* 49-71.

Falk, J. J. H., & Dierking, L. L. D. (2002). *Lessons without limit: How free-choice learning is transforming education*. New York: Rowmand & Littlefield.

Friedman, A. (Ed). (2008). *Framework for evaluating impacts of informal science education projects.* Arlington, VA: National Science Foundation. (Available at: http://caise.insci.org/uploads/docs/Eval_Framework.pdf)

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & gaming*, *33*(4), 441-467.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, *1*(1), 20-20.

Ifenthaler, D., Eseryel, D., & Ge, X. (2012). *Assessment for game-based learning* (pp. 1-8). Springer New York.

Klehe, U. C., & Anderson, N. (2007). Working hard and working smart: motivation and ability during typical and maximum performance. *Journal of Applied Psychology*, *92*(4), 978.

Mayo, M. J. (2009). Video games: a route to large-scale STEM education?.*Science*, *323*(5910), 79-82.

Mislevy, R.J. (2011). Evidence-centered design for simulation-based assessment (CRESST report 800). Los Angeles, CA: The National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles. (Available at: https://www.cse.ucla.edu/products/reports/R800.pdf)

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits.* Committee on Learning Science in Informal Environments. Philip Bell, Bruce Lewenstein, Andrew W. Shouse, & Michael A. Feder, Eds. Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*(3), 482.

Schwartz, D.L. & Arena, D. (2009). Choice-based assessments for the digital age. *MacArthur 21st Century Learning and Assessment Project*.

Schwartz, D.L. & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age.* Cambridge, MA: MIT Press.

Schwartz, D.L., Bransford, J.D., & Sear, D. (2005). Efficiency and innovation in transfer. In J.P. Mestre (Ed.) *Transfer of learning from a modern multidisciplinary perspective*, (pp. 1-51). Greenwich, CT: IAP.

Semmens, R., Blair, K.P., & Schwartz, D.L. (2015). How sick is that doggie in the window? Game choices correlate to academic performance. Manuscript in preparation.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503-524.

Shute, V. J., D'Mello, S. K., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. Computers & Education, 86, 224-235.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In Ritterfeld, U., Cody, M., & Vorderer, P. (Eds.). *Serious games: Mechanisms and effects*, (pp. 295–321). New York: Routledge.

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE Review*, *41*(2), 16.

Tables

Table 1. Examples of curricular standards highlighting data literacy skills

Standard	Domain	Description of student competency
Common Core	Math Practice: Model with mathematics (Math.Practice.MP4)	[Proficient students] are able to identify important quantities in a practical situation and map their relationships using such tools as diagrams, two-way tables, graphs, flowcharts and formulas.
	English Language Arts: Reading science and technical subjects (ELA-Literacy.RST.9-10.7)	Translate quantitative or technical information expressed in words into a visual form (e.g. a table or chart) and translate information expressed visually or mathematically (e.g. in an equation) into words.
NGSS Practices	4. Analyzing and interpreting data	Construct, analyze, and/or interpret graphical displays of data and/or large data sets to identify linear and nonlinear relationships.
	5. Use mathematical and computational thinking	Use mathematical, computational, and/or algorithmic representations of phenomena or design solutions to describe and/or support claims and/or explanations.
	8. Obtaining, evaluating, and communicating information	Critically read scientific literature to determine the central ideas or conclusions; Compare, integrate and evaluate sources of information presented in different formats (e.g. visually, quantitatively), as well as in words in order to address a scientific question or solve a problem.

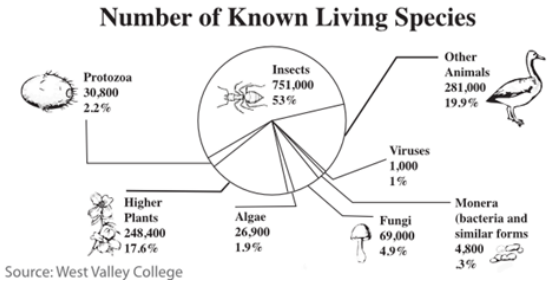
Table 2. School demographics

Demographic Category	% of Students
Hispanic/Latino	58.6
Caucasian	29.4
Asian/Filipino	4.9
Black	2.3
Other/NR	20.8
English language learners	20.8
Socio-economically disadvantaged	37.6

Table 3. Assessment items on Storylet post-test

Item Type	Question
Content focus: multiple-choice	**1. How long do soda bottles take to decompose?** a) 5 years b) 100 yrs c) 1000 yrs d) 20000 yrs
	2. What percent of teens have access to the internet or cell phones? a) 40% b) 55% c) 75% d) 95%
	3. How did Argentinian ants get to California? a) Getting picked up by migrating birds and other animals b) Stowing away on cargo ships c) Escaping from ant farms, which became popular in the 1900s d) They are native to California, so they didn't have to travel there
Data literacy and visualization focus: open-ended	**4. A student named George made this graphic.** *(lower left panel)* **Please write feedback for George to help him make it better.**
	5. Did you choose this graphic? *(lower right panel)* **Why or why not?**

4.



5.

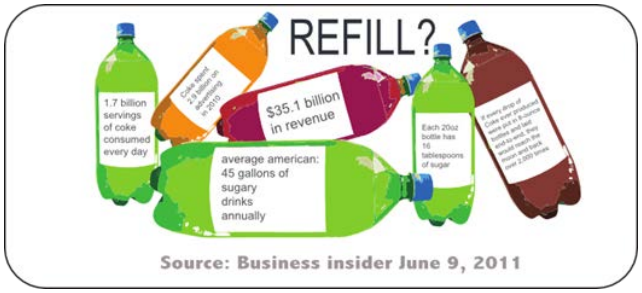


Table 4. Correlations between learning measures (N=87)

	Post-test: Graphics	Total Buzz Meter Score	Buzz Meter: Chart	Buzz Meter: Photo	Buzz Meter: Title
Post-test: Content	.087	.085	-.08	.183	.213*
Post-test: Graphics	-	.384**	.349**	.245*	.235*

Note: * = p < .05 (2-tailed) and ** = p < .01 (2-tailed)

Figure legends

Figure 1. Sample classroom materials: contrasting cases of graphs showing anchovy and sardine catch on different time scales. Students analyzed and compared the various features and main message for each graph, both individually and in small group discussions.

Figure 2. Flowchart of Storylet game mechanics showing one cycle. The game consists of three story cycles in which the player must choose the most appropriate photo, chart, and title for each story.

Figure 3. Chart choices for *Argentine Ant Attacks* story. a) Spread of Argentinian ant across the globe. b) Number of known living species: Insects represent 53% of species. c) Chemical weapon use per experimental fight with Argentine ants by % of winter ants represented. d) Chemical weapon use of winter ants by distance from nest.

Figure 4. a) Mean visits to game tabs, Graphics (charts + photos) vs. Stories. b) Mean total score on Graphics-focused (items 4 - 5) vs. Content-focused (items 1- 3) portions of the post-test. Error bars represent SE.

Figures

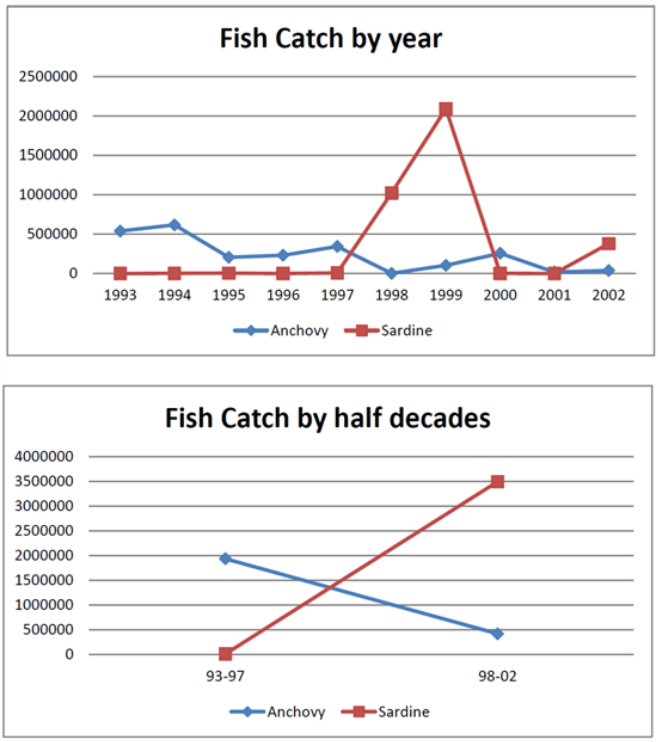


Figure 1. Sample classroom materials: contrasting cases of graphs showing anchovy and sardine catch on different time scales. Students analyzed and compared the various features and main message for each graph, both individually and in small group discussions.

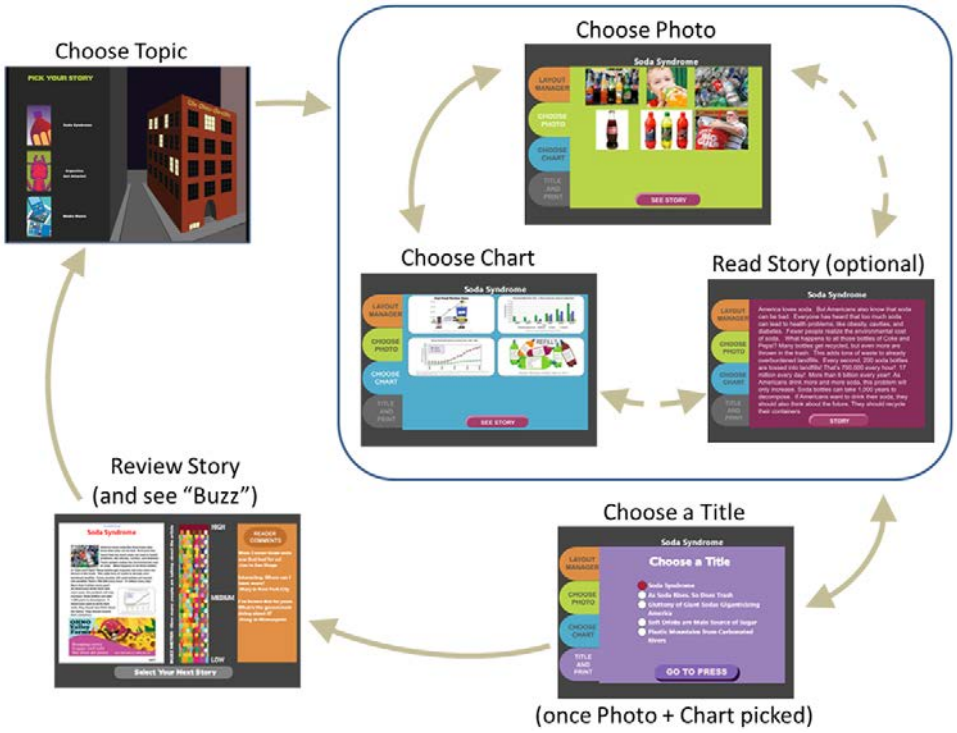


Figure 2. Flowchart of Storylet game mechanics showing one cycle. The game consists of three story cycles in which the player must choose the most appropriate photo, chart, and title for each story.

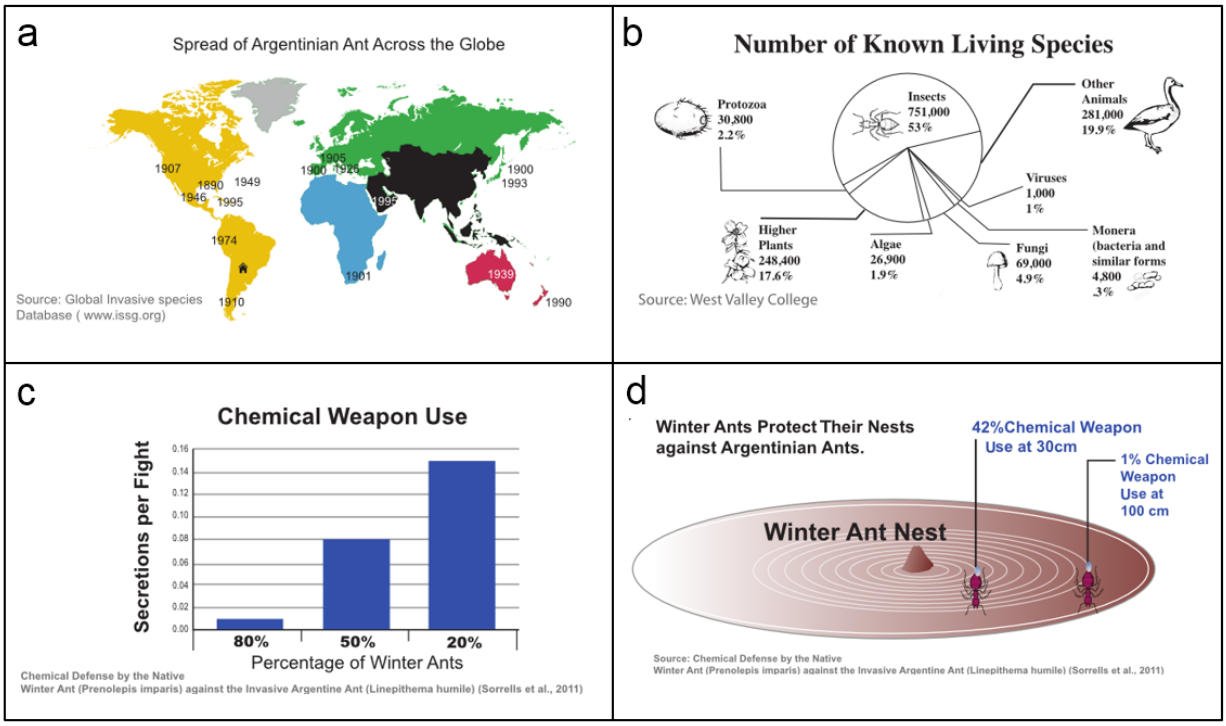


Figure 3. Chart choices for *Argentine Ant Attacks* story. a) Spread of Argentinian ant across the globe. b) Number of known living species: Insects represent 53% of species. c) Chemical weapon use per experimental fight with Argentine ants by % of winter ants represented. d) Chemical weapon use of winter ants by distance from nest.

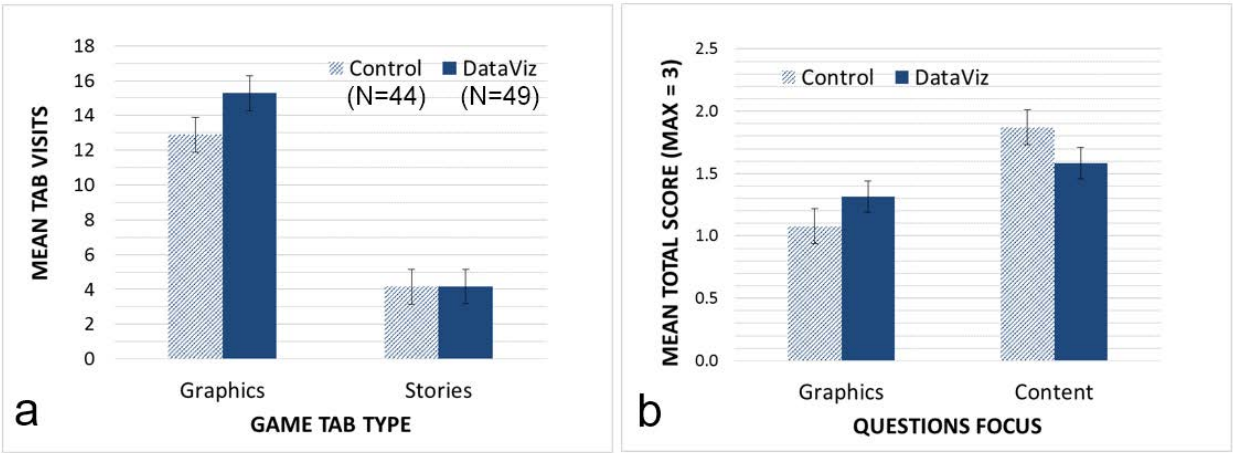


Figure 4. a) Mean visits to game tabs, Graphics (charts + photos) vs. Stories. b) Mean total score on Graphics-focused (items 4 - 5) vs. Content-focused (items 1- 3) portions of the post-test. Error bars represent SE.