International Journal of Artificial Intelligence in Education Volume# (2016) Number
IOS Press

Assessing Whether Students Seek Constructive Criticism: The Design of an Automated Feedback System for a Graphic Design Task

Maria Cutumisu[1], Kristen P. Blair[2], Doris B. Chin[2], Daniel L. Schwartz[2]

Abstract. We introduce a choice-based assessment strategy that measures students' choices to seek constructive feedback and to revise their work. We present the feedback system of a game we designed to assess whether students choose positive or negative feedback and choose to revise their posters in the context of a poster design task, where they learn graphic design principles from feedback. We then describe an empirical study that sampled one hundred and six students from a US middle school to evaluate the feedback system. We make the following contributions: (1) describe the design and implementation of a novel feedback system embedded in an assessment game, Posterlet, (2) outline an approach to analyze graphic design principles automatically to provide contextual feedback in a novel poster design domain, (3) show that choices to seek negative feedback and to revise correlate with in-game performance, and most importantly, (4) show that choices correlate with in-school achievement: the choice to revise correlated with both in-school performance measures (Science and Mathematics grades), while the choice to seek negative feedback correlated with students' prior standardized scores in Mathematics.

Keywords. choice, feedback, performance, learning, assessment, game

INTRODUCTION

The main purpose of formative assessments is to promote learning rather than just "accountability, ranking, or competence" (Black & Wiliam, 2004; 2009). While the literature shows strong evidence of an increase in students' performance when formative assessments are employed (Black & Wiliam, 1998), these assessments are not always implemented in the classroom. The pervasiveness of massive open online courses (MOOCs) calls for scalable formative assessments that can provide immediate, dynamic, and responsive feedback customized to the user and to the learning domain and context. Most learning

[1] Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, AB, Canada, cutumisu@ualberta.ca

[2] Graduate School of Education, Stanford University, Stanford, CA, USA

environments that support feedback are designed for structured domains, such as mathematics, and provide little agency to the users regarding when and how to receive feedback.

Feedback provides information related to a person's performance or understanding (Hattie & Timperley, 2007) and, thus, it is an important, though often neglected, component of the learning process. Consider that most innovations introduced in schools (e.g., computer-assisted instruction, peer tutoring) improve achievement by about 0.4 of a standard deviation (Hattie, 1999). In contrast, twelve meta-analyses focusing on feedback in classrooms, which included 196 studies and 6,972 effect sizes (Hattie, 1999; Hattie & Timperley, 2007), found that the average effect of feedback was nearly double (0.79 SD), surpassing the effects of students' prior cognitive ability (0.71) and homework (0.41). Concomitantly, Kluger and DeNisi (1996; 1998) found in their meta-analysis comprising 131 research studies that feedback was *detrimental* to learning performance in a *third* of the studies. The feedback literature distinguishes between the effects of positive *versus* negative (i.e., critical) feedback. When it is accompanied by corrective information, negative feedback may reduce the discrepancy between understanding and response, in which case it can help people identify skills they should adopt or improve (Dweck & Leggett, 1998; Nussbaum & Dweck, 2008; Trope & Neter, 1994). It also tends to be more effective for continued performance than positive feedback, presumably because positive feedback indicates that one has done enough, whereas negative feedback indicates the need for a change (Hattie & Timperley, 2007). However, critical feedback runs the risk of negative affect and it can trigger ego threat (Kluger & DeNisi, 1998).

The feedback literature was launched from the behaviorist tradition (Thorndike, 1927). Thus, most studies focus on supervised feedback, where it is up to the teacher, experimenter, or computer to decide when and how to deliver feedback. However, in many situations, people need to actively seek feedback. In organizational studies, research on feedback seeking has focused mainly on work performance outcomes (Ashford & Cummings, 1983; Anseel et al., 2013). Little is known about the implications of students' feedback choices on their learning or about variables that influence students' feedback choices, but researchers acknowledge the importance of the mechanisms underlying feedback for learning. We view feedback choice as an important construct for self-regulated learning. Feedback that is informative and actionable for learners is a crucial factor in developing mastery (Ericsson, Krampe, & Tesch-Römer, 1993; Schwartz, Tsang, & Blair, 2016), especially when it is provided immediately (Kulik & Kulik, 1987). Zimmerman (1990) included *responsiveness to self-oriented feedback* among three critical features of students' self-regulated learning strategies. For example, revision is an important activity often triggered by critical feedback.

In this manuscript, we focus on students' choices to seek negative feedback and to revise, and we describe Posterlet, a game-based assessment instrument we developed to measure these choices. Posterlet was designed following the principles of choice-based assessments (Schwartz & Arena, 2013) that focus on assessing students' learning processes, rather than only their learning outcomes. The underlying principles include: (1) preparation for future learning (PFL; Bransford & Schwartz, 1999) - provide students with learning opportunities during the assessment, (2) free choice – provide students with multiple pathways to complete the assessment regardless of their choices, and (3) typical performance – provide students with the conditions in which they can perform their natural, rather than maximal or test, behaviors (Klehe & Anderson, 2007). Game-based assessments, such as Posterlet, collect process data (e.g., the choices students make while learning in new contexts) instead of relying on self-report measures and pen-and-paper tests. Our choice-based assessment approach is similar to other computer-based learning technologies that collect and analyze process data. For example, domain-specific intelligent tutoring systems (ITSs; Corbett & Anderson, 2001; Koedinger, Anderson, Hadley, & Mark, 1997) collect measurements mainly to determine ways to help students learn more domain-specific cognitive skills from the system, which is an important goal. However, our aim is to examine whether various forms of self-regulated learning (SRL) instruction prepare students to learn in new contexts. For example, in a different study, we taught design thinking skills (e.g., seeking negative feedback from peers and instructors) to middle-school students for five weeks as part of their regular classroom instruction (Conlin, Chin, Blair, Cutumisu, & Schwartz, 2015). We integrated the instruction into their existing curriculum via several subject areas (Math, Social Studies, and Science). At the end of the five-week classroom-based instruction, we employed Posterlet as a post-test to assess how well students learned to seek negative feedback. This research presents an approach to assessment that examines key student choices while learning, rather than only the outcomes of learning.

Much research has already been conducted in domain-general metacognitive tutoring systems to assess metacognitive tutoring, its impact on learning outcomes, and its transfer capacity (Aleven & Koedinger, 2000; Clark & Mayer, 2016; Conati & Vanlehn, 2000; Roll, Aleven, McLaren, & Koedinger, 2011; Roll, Baker, Aleven, & Koedinger, 2014). Similar to Posterlet, these approaches have empirically shown examples of metacognitive tutoring strategies based on students' domain-general choices that positively impacted students' learning outcomes within the tutoring system, outside of the tutoring system, and transferred to other contexts. Moreover, the feedback system that stands at the core of the Posterlet assessment game draws from the constraint-based tutoring architecture (Mitrovic & Ohlsson, 2016). Specifically, the domain model is capable of evaluating a given poster and providing feedback to the students, without providing hints on how to improve performance.

Although our approach is similar to that of learner-control domain-general metacognitive tutoring systems, there are several main distinctions. First, the Posterlet game is not designed to be a tutoring system. Specifically, the game does not teach learning strategies, such as choosing negative feedback or choosing to revise. Thus, we do not expect students to choose negative feedback more or to choose to revise more by the end of the game, and we did not find this effect in our studies. Second, the purpose of the game is not to teach domain knowledge skills (e.g., graphic design principles for poster design). Players are not guided through tutorials about graphic design principles, but they have at most two chances on each poster (i.e., the second feedback choice of the same valence is always uninformative) to find out information about their use of a graphic design rule (i.e., feedback) and not about how to use that rule (i.e., help). Instead, Posterlet is designed as an instrument that assesses how often students choose to seek negative feedback and to revise. Lastly, Posterlet targets a novel open-ended domain, poster design.

We have previously described the Posterlet assessment system, which enables us to measure the influence of feedback and revision choices on learning in an open-ended environment. In this article, we focus particularly on the design of the underlying feedback system that automatically evaluates posters for the use of 21 graphic design principles and subsequently generates a pool of positive and negative feedback. Players have an opportunity to choose either positive or negative feedback from virtual characters, as well as to choose to revise their current poster after reading the feedback for that poster. It is important to note that (a) both types of feedback are designed to be equally informative and (b) negative feedback here denotes constructive criticism and not punishment. Prior studies of Posterlet have shown that seeking negative feedback and revising are good choices for learning in the context of the graphic design domain. We found evidence that seeking negative feedback and revising in the game was associated with enhanced performance (as measured by final poster quality), as well as improved learning (as measured by a follow-up test of graphic design principles; Cutumisu, Blair, Chin, & Schwartz, 2015).

Replicating some of our prior findings, we present previously unreported data examining the relationship between in-game choices, in-game performance, and school achievement. Specifically, we conducted a study to address the following research questions:

1) Do students' feedback and revision choices relate to learning in the assessment environment?

2) Do students' feedback and revision choices relate to outside measures of school achievement?

POSTERLET: A GAME TO ASSESS STUDENTS' LEARNING CHOICES IN A POSTER DESIGN CONTEXT

In the Posterlet game, the player is a member of the organizing committee for the school's Fun Fair with the task of designing posters for three activity booths (see Figure 1 for game flow).

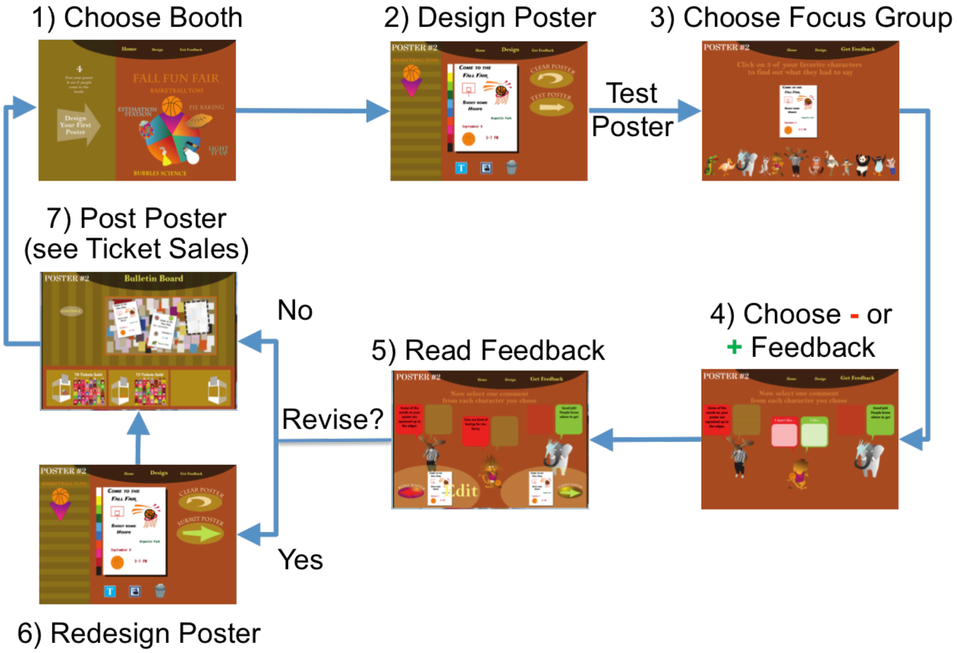
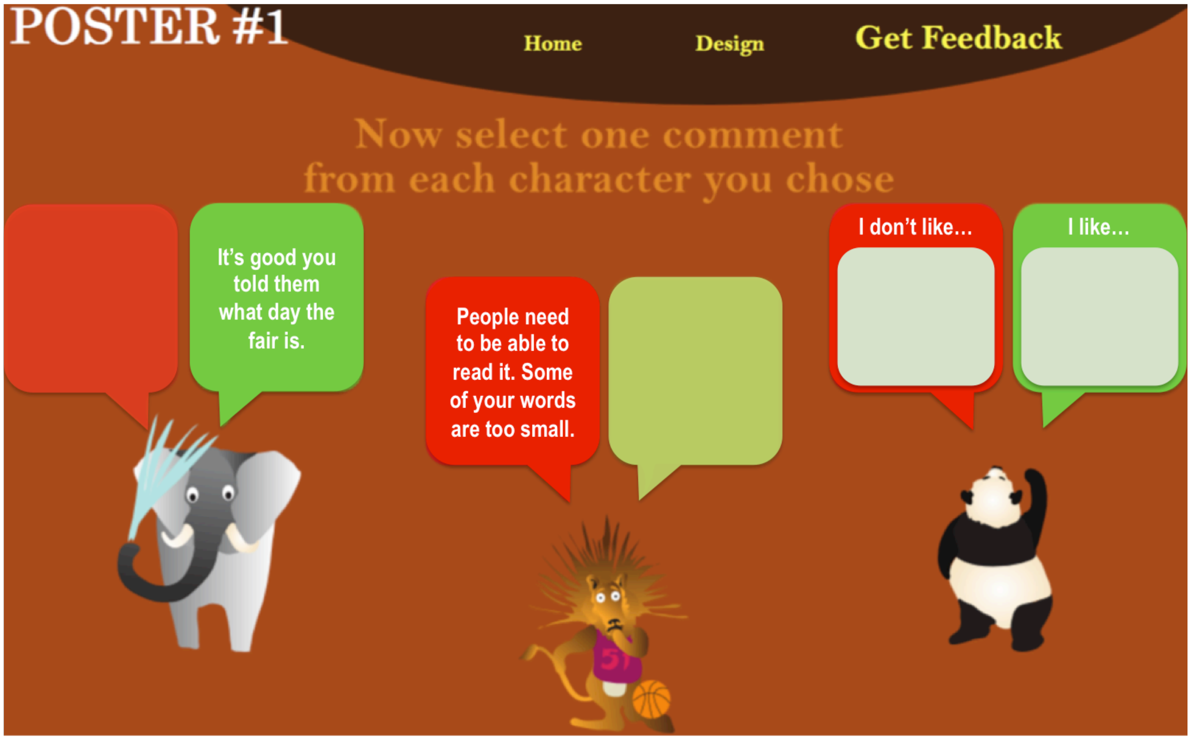


Figure 1. Posterlet game design flow. (Reprinted from Cutumisu, Blair, Chin, and Schwartz, 2015.)

First, the player chooses a theme (e.g., basketball toss, science of bubbles, estimation station, pie-baking, and “light it up” electricity). Next, the player designs the poster, adding text phrases and graphics from a pre-populated text and graphics palette, respectively, on the canvas. The player may adjust the text and background colors, the size of the text boxes and graphics, the font’s size, style, and alignment, or the location of the items. The text palette provides a set of phrases that are common to all themes (e.g., the time and place of the event) or that are specific to each booth theme (e.g., the booth descriptions). Similarly, the graphics palette provides a set of five graphics for each booth theme (e.g., a hoop for the basketball toss theme). When the player indicates that the poster is ready by pressing the “Test” button in Step 2, twelve animal characters gather to view the poster and form their opinions about that poster. Then, the player selects three of these characters and asks each of them to provide feedback. Each character displays two boxes: one that corresponds to positive feedback (“I like…”) and one that corresponds to negative feedback (“I don’t like…”), as illustrated by Step 4 and Figure 2.



Now select one comment from each character you chose

It’s good you told them what day the fair is.

People need to be able to read it. Some of your words are too small.

Figure 2. Students may choose either positive or negative feedback from each character. Here, a student chose one positive feedback from the elephant, one negative feedback from the lion, and nothing yet from the panda. (Reprinted from Cutumisu, Blair, Chin, and Schwartz, 2015.)

The player may only click on one of these two boxes per character, choosing either positive or negative feedback from each character. For example, when choosing positive feedback, the player may receive the following message: “It is good that the text is large enough to read.” The equivalent message when the user chooses negative feedback is the following: “It is hard to read. The text is too small.” Then, the player reads the feedback, as shown in Step 5. It is important to note that positive and negative feedback messages were deliberately designed to match for length and informational value to avoid confounding the message valence with the message content. Finally, the player chooses to revise or submit the poster. There is only one round of feedback and revision for each poster; the player is not offered feedback after revising a poster. Upon poster submission, the player sees the poster’s score (i.e., number of tickets sold for that booth), as shown in Step 7. The computation of the score is detailed in the *Method* section. The game has three rounds, with a total of nine feedback choices and three revision choices.

FEEDBACK SYSTEM’S GRAPHIC ANALYSIS: DESIGN AND IMPLEMENTATION

We designed our feedback system as a visual graphic analyzer that evaluates an open-ended design domain and creates customized feedback for learners. The goal of the system is two-fold: 1) to generate a feedback message based on the content of a poster, on the feedback valence (positive or negative) chosen by the player, and on the player’s history throughout the game, and (2) to collect data regarding players’ choices and task performance The feedback system evaluates each poster and generates a pool of eligible feedback messages. These messages are ordered by a priority scheme and distributed to players according to their choices for positive or negative feedback.

Feedback Generation

When the player submits a poster, the feedback system evaluates that poster on 21 dimensions of interest (see Table 1). These features represent simplified rules of graphic design gathered from a consulting professional graphic artist.

Table 1. The 21 graphic design principles evaluated for each poster by our system for feedback generation. (Adapted from Cutumisu, Blair, Chin, and Schwartz, 2015.)

Information	**Readability**	**Space Use**
location present	text within bounding box	space used by graphics within range
date given	text size large enough	top half used
time given	font style readable	bottom half used
ticket price given	text & graphics disjoint	graphics spaced away from edge
booth description present	text & text disjoint	text spaced away from edge
graphics relevant	graphics size large enough	text spaced away from other text
text present	text contrast high	--
graphics present	--	--

We grouped these composition rules into three categories: (1) *Relevant Information*: important details of the event promoted by the poster (e.g., date, time, and location), (2) *Readability*: visibility of poster text and graphics (e.g., color contrast between the text and the background, font and graphics size), and (3) *Space Use*: use of space on the poster canvas (e.g., distance between the text and the edge of the poster).

There is an additional fourth category, called uninformative feedback, which does not provide any task-related information to the player (e.g., positive: “I like fairs” or negative: “I don’t really like fairs”).

The system starts by checking the correctness of the poster’s entities (i.e., text boxes and graphics) in the order in which they were placed on the canvas. For each entity, the system checks all the rules applicable to that entity and determines whether each rule was used correctly, incorrectly, or not used at all. For instance, to analyze a rule about color contrast between the text and the background, we implemented an algorithm that computed contrast and relative luminance (Luminance Algorithm, 2016). The same rule could be applied both correctly and incorrectly on a poster, for different entities. For example, a player could add two text boxes with different text colors on a poster, one text box creating a low contrast and the other creating a high contrast with the poster’s background color. This would trigger an incorrect use and a correct use, respectively, of the color contrast rule. Some rules check a feature in isolation (e.g., location present) and return a binary result. Others take into account the rest of the entities on the poster (e.g., overlap between text and graphics). For instance, two text boxes placed side by side that touch, but do not overlap, trigger an incorrect use of a “text spaced away from other text” rule, only if the text of the left text box is aligned to the right and the text of the right text box is aligned to the left. Finally, rules such as “graphics relevant to the poster’s theme” rely on the booth’s theme and a list of pre-populated relevant graphics for each theme maintained by the system, so that a soccer ball used on a basketball poster would trigger an incorrect use of this rule. As a result of evaluating the graphic design principles for a poster, the system builds an array of rule objects. Each rule object contains a feature (e.g., text size), the number of correct uses of that feature, the number of incorrect uses of that feature, the feature’s order within the feedback set presented to the player, a positive feedback message, and a negative feedback message.

Feedback Prioritization

We prioritized the feedback features to ensure a) balanced coverage of the design features, b) non-repetition of feedback, and c) alternation of informative and uninformative feedback within a feedback valence to avoid cognitive overload of feedback information for younger participants (see Figure 3).

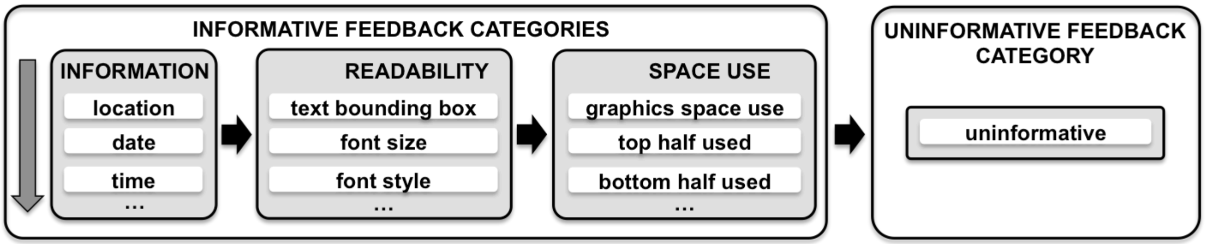


Figure 3. Posterlet’s feedback priority: Informative Categories are prioritized from left to right, alternating with the Uninformative Category. Features are prioritized from top to bottom within a category.

Balanced Coverage. To ensure that the features covered all categories equally, we first prioritized the three informative feature categories from highest (Relevant Information) to medium (Readability) to lowest (Space Use) priority. Second, we prioritized the features within each category (e.g., “location” has a higher priority than “date” in the Relevant Information category) to ensure that the highest priority feature is always selected from each category. Every time the player chooses feedback, the system first iterates through the three informative feedback categories, followed by the fourth uninformative category, searching for the first eligible feature with the highest priority (shown in Figure 3), based on the array of rules extracted from the poster upon its submission. Initially, this category is Relevant Information. If a suitable feature is found in this category, the feature is retrieved, the search ends, and a feedback message is displayed to the player (detailed in the *Feedback Message Retrieval* section). A subsequent search continues in the next informative category. If no feature is eligible in any of the three informative categories (e.g., when all eligible features have been exhausted from all informative categories), the system selects a feedback message from the Uninformative Feedback Category, shown in Figure 3. The pseudo-code of the feature-retrieval technique and of the algorithm we designed and implemented in JavaScript for our feedback system are presented in Table **2** and Table **3**.

Table 2. Find an eligible feedback feature for the feedback valence chosen by the player

```
// Return the highest priority feature available for the valence chosen by the player.
findFeedbackFeature (rules, valence):

  // Default message: uninformative
  feature ← uninformative
  // Iterate through all four categories
  for category in categories
    // Select a feature for a category and feedback valence
    for feat in category
        if feat unused and applicable for valence
           turn feat off
           feature ← feat
           return feature
        // end if
    // end for
  // end for
  return feature
```

Table 3. The algorithm underlying the feedback system

```
Algorithm:

// 1: Build the rules array
// When the player submits a poster, parse that poster and update the array of rule objects.
rules ← parsePoster(poster)
```

```
// 2: Find an eligible feedback feature based on the poster rules and chosen feedback valence.
feature ← findFeedbackFeature (rules, valence)

// 3: Retrieve a feedback message for the player
// Generate a feedback message for an eligible feature and the chosen feedback valence.
 message ← retrieveFeedbackMessage(rules, feature, valence)
```

Non-repetition. Players do not receive feedback on the same feature twice, regardless of the feedback valence they chose. The system manages the eligibility of the features by employing an array that keeps track of the status (on or off) of each feature as the player selects feedback. For example, when the player chooses positive feedback, the system retrieves the first eligible feature (i.e., used correctly by the player on that poster and that was not presented to that player before). A feedback feature presented to a player is turned off subsequently. One of the most important conditions for feedback effectiveness is the extent to which feedback is easy to understand and to relate to the student's previous knowledge (Hattie & Timperley, 2007). Creating customized feedback that takes into account the student's performance across all the posters aims to satisfy this condition. Additionally, we wanted to avoid a scenario in which a virtual character provided the same feedback on the same feature to the same player twice. This principle aims to maximize the number of different graphic design rules a player would encounter and to provide a more natural interaction of the player with the virtual characters. For example, in most online games, meeting the same virtual characters again prompts them to change their greetings or behaviors to create the illusion of a rapport with the player.

Alternation of Informative versus Uninformative Feedback. To address a potential information overload for some of the younger players, the priority scheme also ensures that feedback of the same valence alternates between informative and uninformative. For each poster, the first feedback message of a given valence is always informative and the second feedback message of the same valence is always uninformative. This ensures that both positive and negative feedback have equal chances to be informative. The system keeps track of the feedback value (informative or uninformative) shown to the player locally (for each poster), as well as globally (across posters). Thus, an index to the next informative category in the sequence (Relevant Information, Readability, or Space Use) is retained across posters, so that on the next iteration the search starts in that category. The assumption behind this design decision is that players may infer rules from the same category (*day* of the fair) upon receiving feedback from that category (*time* of the fair). Cycling through the three categories of feedback first, before descending into each category, increases students' chances of learning as many rules from the feedback as possible.

Feedback Retrieval

As discussed above, the system evaluates each poster and generates an array of rule objects. Each of the design features has a corresponding rule object that contains a pool of both positive and negative feedback messages. Since informative feedback features are not repeated for a player, only one message per feedback valence suffices. In contrast, uninformative feedback consists of several options to create a more realistic experience for the player, because uninformative feedback appears many times across posters. For example, uninformative feedback phrases are selected from nine different positive phrases (e.g., "Yay!

Fairs are fun!", "I always go to fairs.", etc.) and nine different negative phrases (e.g., "Hmm, fairs are boring.", "I never go to fairs.", etc.).

When a player submits a poster draft, the system first parses the poster according to the 21 graphic design principles into an array of rules, each accompanied by its frequency of correct and incorrect uses (Step 1 in Table **3**). Then, when the player chooses feedback, the feedback system selects a suitable feature based on the array of rules by: 1) retrieving the next feature in the priority order for which feedback has not been provided already, according to the non-repetition principle (Step 2 in Table **3**), 2) examining the frequency of use for the feedback valence chosen by the player to decide whether the feedback is informative or uninformative, and 3) retrieving and displaying the feedback message to the player (Step 3 in Table **3**). When the priority scheme determines that a particular informative feature is next in line, its positive feedback message can only be deployed if the array indicates a correct usage of that feature. Similarly, a feature's negative feedback message is only eligible for retrieval if the array indicates an incorrect usage. When informative feedback cannot be generated, the system defaults to the uninformative feedback category. For instance, when requesting negative feedback, if the player has not broken any rules or has already received negative feedback for all incorrectly used features (feedback for any particular rule is not repeated, according to our *Non-repetition* principle), the player will receive one of the many uninformative negative feedback messages (e.g., "I don't like fairs.").

METHOD

Participants

One hundred and six grade 8 middle-school students (60 female, 46 male), aged 13 to 14 years, from a public middle school in California, participated in a study involving Posterlet in May 2015. All students had the same science teacher and the most recent school accountability report card indicates student enrollment of 591 5th – 8th graders (56.5% White, 18.1% Hispanic/Latino, 11.2 % Asian/Filipino, 12.9% two or more races, 1.4% Other; 8.6% socioeconomically disadvantaged and 5.4% English language learners). Table **4** provides a summary of the school and participants' information, with income data retrieved from http://factfinder2.census.gov based on the school's location.

Table 4. School and participant information

			Posterlet (n = 106)	
Median Family Income	**Grade**	**Age Range**	**Females**	**Males**
$143,889	8th	13 - 14	60	46

Procedure, Materials, and Design

The Posterlet game was included as one of a series of activities in an unrelated study. The procedure was the same for all students: participants arrived in their science classroom and were informed about the activities unfolding during their class period. Students logged on and played the game individually for an average of M = 14.8 minutes (SD = 4.07). Eight students did not complete the Posterlet game due to time considerations, so we excluded these students from our analyses. We also excluded nine students for whom we did not receive consent. Thus, we included n = 89 students (50 females, 39 males) in the analyses. Some parents did not provide consent to sharing their children's grades or standardized test scores, so we removed students from the data set as needed, depending on the analyses we conducted. For example, we excluded all students whose parents did not consent to disclosing school grades from the analyses that pertained to school grades, but we included them in analyses related to the Posterlet game measures.

Measures

We analyzed choice and learning (performance on the poster task and achievement in school) measures.

Choices. We focused on two measures of students' choices: Feedback and Revision. Negative Feedback (NF) is a whole number representing the total amount of negative feedback chosen by the student across the game and it ranges from zero (i.e., only positive feedback was chosen) to nine (only negative feedback was chosen). Note that participants have control only over their feedback valence (positive or negative), not over their feedback value (informative or uninformative). We also recorded the amount of negative feedback students chose by game round into the following measures: Negative Feedback L1, Negative Feedback L2, and Negative Feedback L3, respectively. Similarly, we recorded the amount of positive feedback students chose by game round into the following measures: Positive Feedback L1, Positive Feedback L2, and Positive Feedback L3, respectively. On each game round, the total amount of negative plus positive feedback chosen equals three.

The second measure of student choice is Revision, defined as the total number of posters a student chose to revise. It is a whole number ranging from zero (i.e., none of the three posters was revised) to three (i.e., all three posters were revised). We collected both types of measurements across the game and each of the three game rounds. The statistics of our measures across the game (Table **5**), as well as the histograms depicted in Figure **4**, Figure **5**, and Figure **6**, indicate non-normal distributions of our measures by game round.

Table 5. Descriptive statistics across game: Negative and Positive Feedback, Revision, and Poster Quality

			Skewness		**Kurtosis**	
Measures (n = 89)	**Mean**	**SD**	**Statistic**	**SE**	**Statistic**	**SE**
Negative Feedback	5.80	2.06	-.55	.25	.02	.51
Positive Feedback	3.20	2.06	.55	.25	.02	.51
Revision	2.02	1.02	-.76	.25	-.54	.51
Poster Quality	36.58	11.48	-1.22	.25	2.38	.51

The negative skewness of Negative Feedback and Revision shown in Table 5 suggests that there are many more scores that are higher rather than lower. The positive kurtosis of Negative Feedback suggests a distribution with a more acute peak around the mean and fatter tails compared to the theoretical normal distribution, while the negative kurtosis of Revision indicates a distribution with a lower, wider peak around the mean and thinner tails. We present descriptive statistics of choices by round and revision status in Figure 7, Figure 8, and Figure 9.

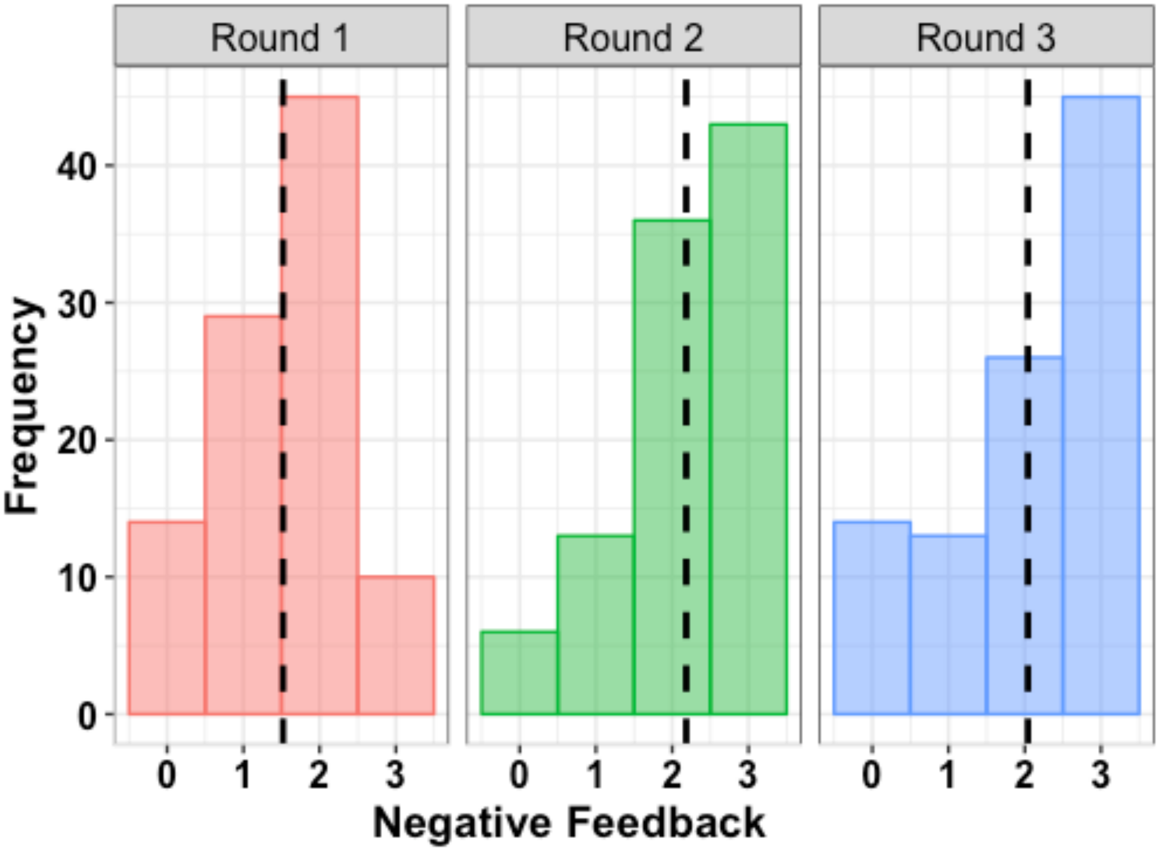


Figure 4. Histograms of students' Negative Feedback choices by game round

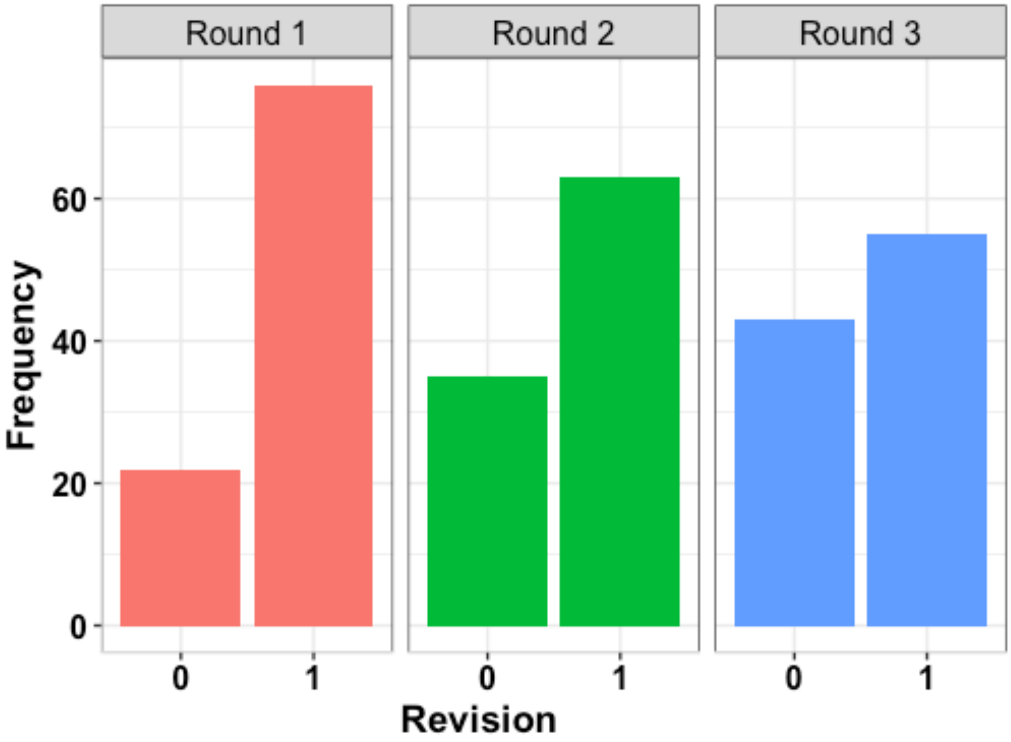


Figure 5. Histograms of students' Revision choices by game round

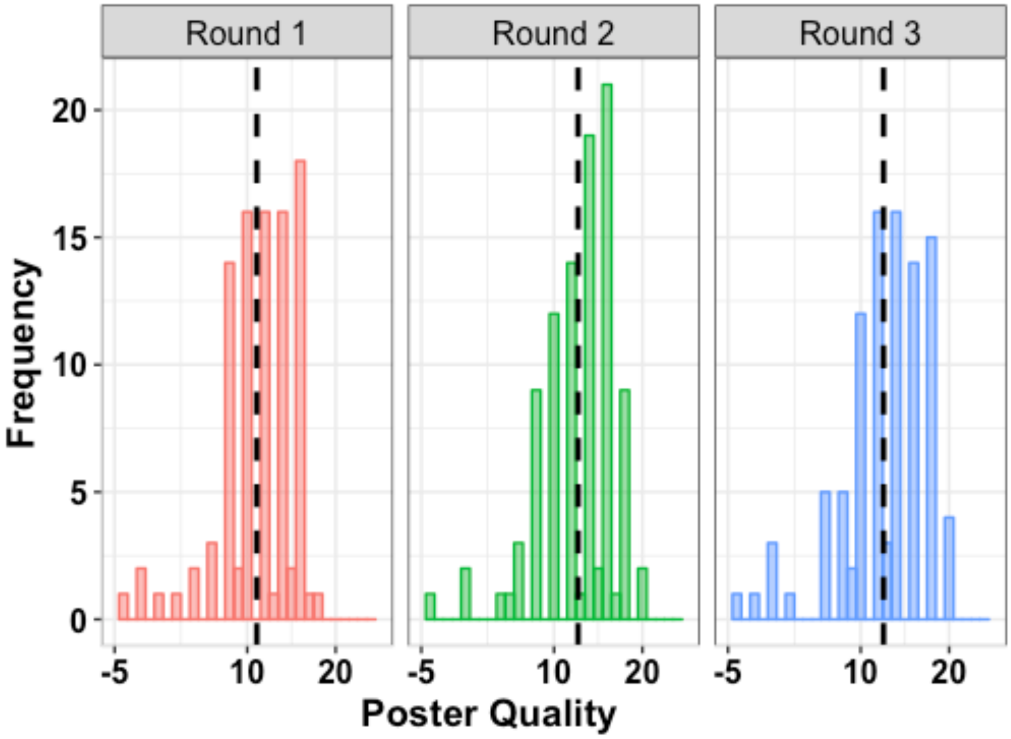


Figure 6. Histogram of students' poster performance, Poster Quality, by game round

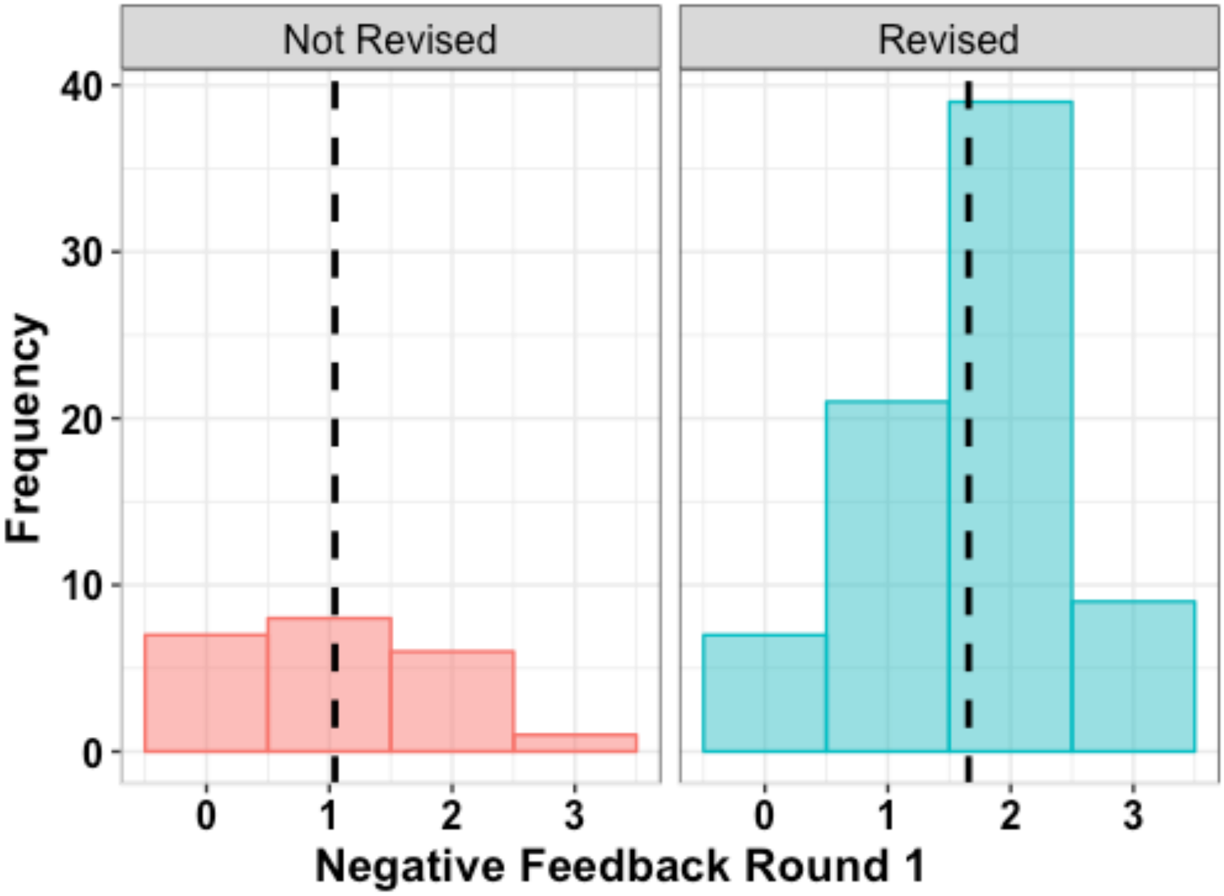


Figure 7. Histograms of students' Negative Feedback choices by revision status on the first game round

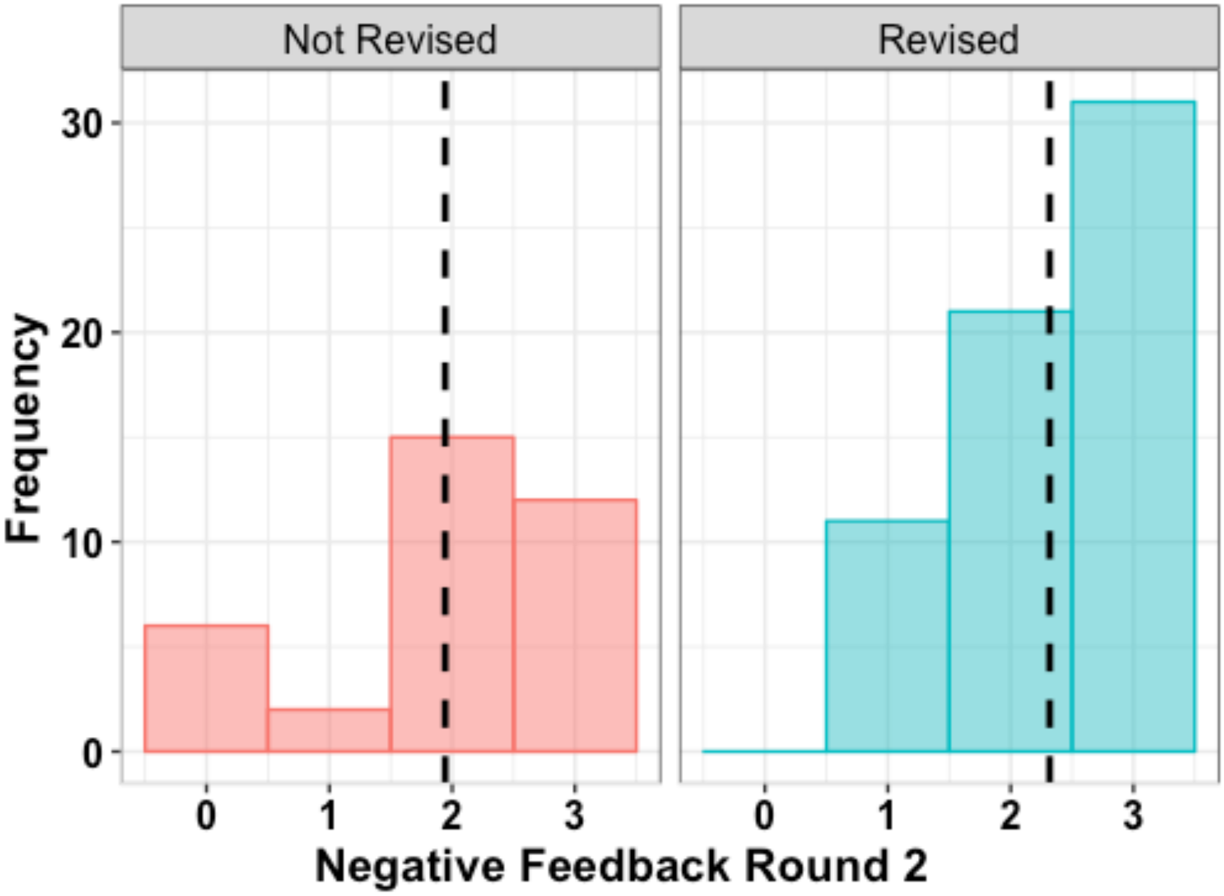


Figure 8. Histograms of students' Negative Feedback choices by revision status on the second game round

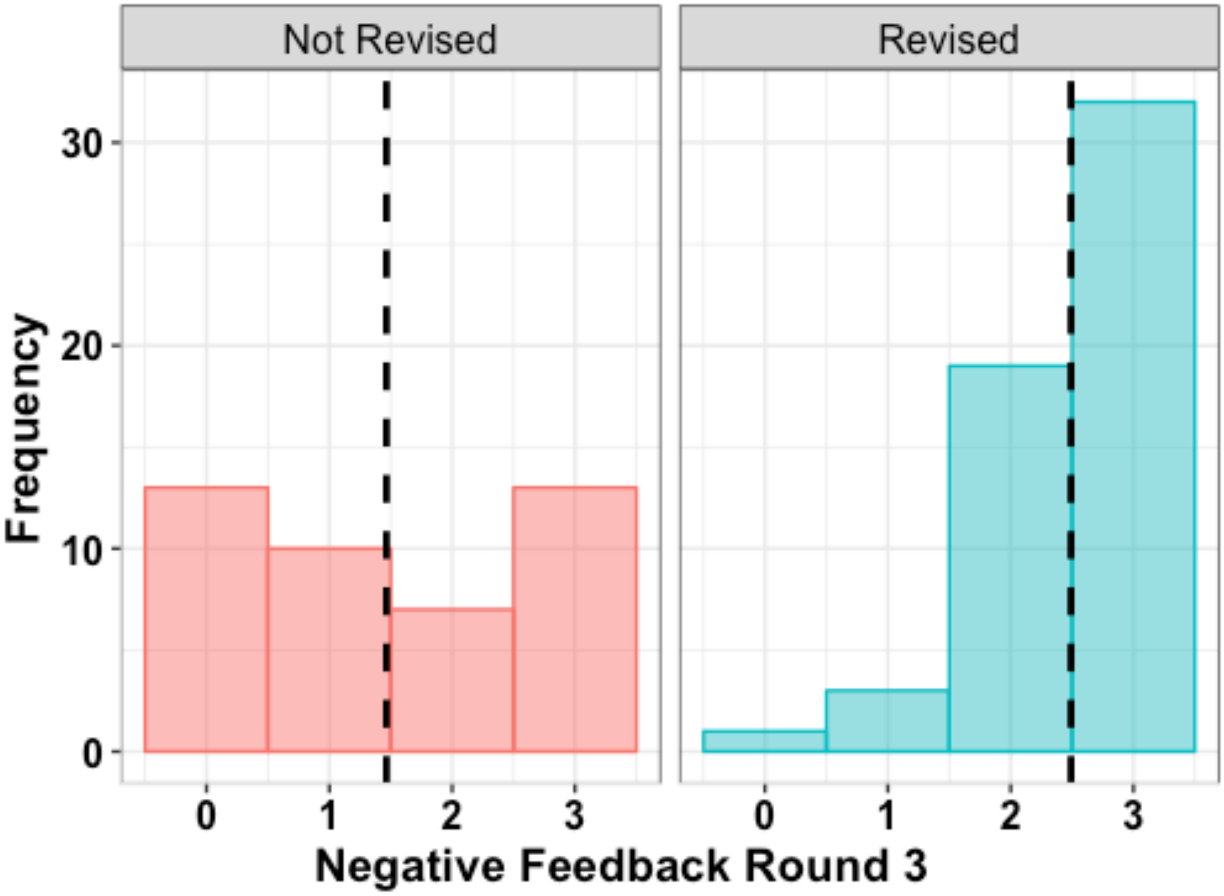


Figure 9. Histograms of students' Negative Feedback choices by revision status on the third game round

Duration. We measured the time students spent on poster design across the game (*Game Duration* in seconds), as well as at each game round. Each round corresponds to posters created for a specific booth (the first poster on that round and a potential revision). We also measured students' time spent reading the feedback across the game (*Feedback Duration* in seconds), as well as on each game round.

In-Game Learning Performance. Poster scores are used as measures of students' in-game performance. They are determined automatically by the feedback system. Each of the 21 graphic design principles could be used correctly, used incorrectly, or not used at all, in different situations on the same poster or across the game. The scoring of a poster keeps track of the number of times a feature is used correctly (*correct count*) and incorrectly (*incorrect count*), on each poster. Thus, we define a tuple (*feature*, *correct count*, *incorrect count*) for each feature. When a feature is not used (i.e., it is not present on the poster or not applicable), the values of *correct count* and *incorrect count* are both zero, whereas when a feature is used either correctly or incorrectly on a poster, one or both of the counts are strictly positive. For example, three different text boxes can use different font styles, two easily readable and one difficult to read. In this case, the system computes a tuple, (font style readable, 2, 1), by increasing the *correct count* of the "font style readable" feature to two and the *incorrect count* of the feature to one. Then, the feedback system scores each feature by assigning:

(a) one point, if the feature is always used correctly but with no incorrect uses, (*feature*, *correct count* > 0, 0). For example, a feature that is used once correctly and never incorrectly, (*feature*, 1, 0), is scored with 1.

(b) minus one point, if the feature is used incorrectly, regardless of any correct uses, (*feature*, *correct count*, *incorrect count*). For example, a feature that is used once correctly and once incorrectly, (*feature*, 1, 1), is scored with -1.

(c) zero points, if the feature is not used on the poster, (*feature*, 0, 0). For example, a graphics size feature on a poster with no graphics is scored with zero.

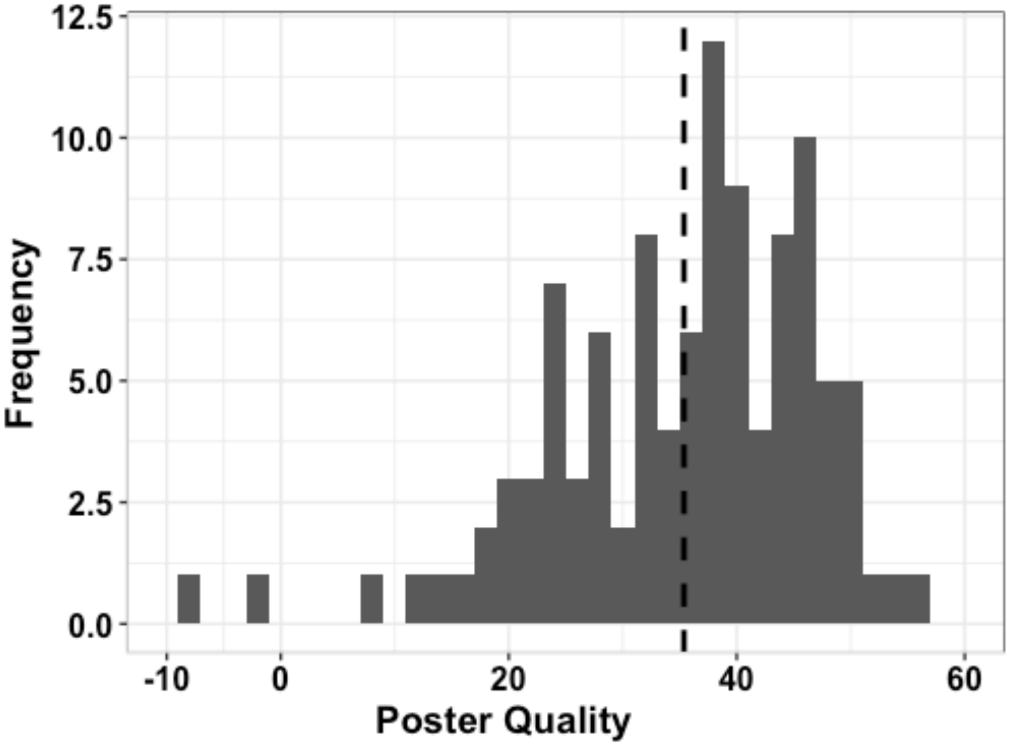


Figure 10. Histogram of students' poster performance, Poster Quality, summed across all game rounds

As a final step, individual feature scores were summed up to create an individual *Poster Score* ranging from -21 to 21. Then, individual poster scores for each of the three rounds were combined to create the overall *Poster Quality* score. The histogram of the Poster Quality measure shown in Figure **10** indicates a non-normal distribution. The negative skewness of Poster Quality shown in Table **5** suggests that there is a larger number of higher than of lower scores. The positive kurtosis of Poster Quality suggests that the distribution has a more acute peak around the mean and fatter tails compared to the theoretical normal distribution.

In-school Performance. Three sets of measures were used as indicators of student achievement outside the game: *Science grades*, *Math grades*, and STAR (Standardized Testing and Reporting) scores (*ELA-CST* and *Math-CST*). We received the average yearly Science grades for all the students (n = 106) and Mathematics grades for a subset of students (n = 65). For the Mathematics grades, we obtained students' grades for the first three trimesters of the current academic year (*Math Trimester 1*, *Math Trimester 2*, and *Math Trimester 3*) and we computed a total average across these three grades (*Math Grades*). Finally, we also employed n = 85 STAR scores in English Language Arts (*ELA-CST*) and Mathematics (*Math-CST*) from two years prior to running our study, administered when these students were in grade 6. These were the last available tests, because California standardized achievement tests were not administered in 2014, due to the transition period to the new Common Core tests.

Data Analyses

Correlations and General Linear Models. We conducted several nonparametric Spearman correlations, as well as regressions and ANOVAs, to assess the degree of association among our measures.

Multilevel Modeling: Developing a Two-Level Hierarchical Linear Model of Individual Change. We built multilevel categorical generalized linear mixed models (GLMMs) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) glmer multilevel analytic routine with a Binomial error distribution in the R Language (R Core Team, 2016) to gain an insight into students' individual revision change based on repeated observations for each student. There are many formal specifications for these models in the literature (Heck, Thomas, & Tabata, 2013; Raudenbush & Bryk, 2002). We examine the change in revision over time (i.e., each game round), within and between students, so we model the probability of revising (Y=1). We focus on the effect of internal (in-game) and external (in-school) performance on revision over *time*. Here, *time* denotes the game round in the Posterlet assessment instrument when revision was measured, being coded as: 0 (round$_1$), 1 (round$_2$), and 2 (round$_3$).

We constructed two-level models, where Level 1 represents repeated measures of revision across the three rounds of the game nested into Level 2, which represents the students. The *intercept* is defined as the natural log of the odds that Y = 1 (e.g., a student is *revising*) when all the predictors X (e.g., *time*) in the model are zero. The intercept is the level of the outcome (e.g., whether the student chose to revise or not) at a point in time (e.g., on round 1). That way, we can compute the likelihood that a student revises on any round of the game. These types of models enable the examination of differences in revision at several time points between groups of individuals.

Results

1. Examining the relation between students' choices and learning in the assessment environment

These analyses are organized into three questions:

a. Do choices to seek negative feedback and to revise correlate with poster performance in our assessment environment?
b. Do choices and in-game learning change over time?
c. Do choices and time spent on feedback affect in-game performance?

1a. Do choices to seek negative feedback and to revise correlate with poster performance in our assessment environment?

Validation of choice is an important step in developing choice-based assessments that focus on the learning process (i.e., students' strategies, such as choosing to seek negative feedback, while solving a problem). If there is a positive association between students' choices and their performance, we can more confidently look at choice as a first-order construct for learning and focus on refining a new wave of choice-based assessments centered on this construct. We can also teach students who struggle to choose helpful strategies that would lead to better performance and learning. In our previous research, we found that the more the students chose to seek negative feedback and to revise, the more they learned from the game. This was evidenced by both their improvement from the initial posters to the final posters and their performance on a post-test of graphic design principles (Cutumisu, Blair, Chin, & Schwartz, 2015).

For this current study, Table **6** shows the Spearman correlations between the non-normally distributed choice (*Negative Feedback* and *Revision*) and poster design performance (*Poster Quality*) measures. Due to time constraints in this study, we were not able to administer an outside post-test of graphic design principles.

Table 6. Spearman correlations between negative feedback, revision, and in-game performance

Measures (n = 89)	Revision	Poster Quality
Negative Feedback	**.40****	**.37****
Revision	--	**.26***

** $p < .01$ * $p < .05$

We found that poster performance correlated moderately with both choices (Cohen, 1988), but more strongly with Negative Feedback. To clarify the contribution of each choice to learning, we conducted a standard linear regression by entering Negative Feedback and Revision together to predict Poster Quality. The prediction model was statistically significant, F(2, 86) = 9.13, $p < .001$, accounting for approximately 17% of the variance of Poster Quality (R Square = .17, Adj. R Square = .16). However, examination of the coefficients reveals that while Negative Feedback was a significant predictor for Poster Quality (Beta = .37, $p = .001$, $t(88) = 3.40$), Revision was not [Beta = .10, $p = .34$, $t(88) = .95$].

1b. Do choices and in-game learning change over time?

We examined how students' in-game choices and learning varied across the game rounds to better understand the choice-performance relation.

First, we focused on the choices and in-game performance over the three rounds of the game. We conducted a repeated-measures analysis of variance to examine the individual change over time of the choices and performance, respectively. We found that performance (Poster Quality) improved across the game: Round 1 = 11.26, Round 2 = 12.71, Round 3 = 12.62; Wilks' Lambda = .91, F(2, 87) = 4.53, p = .013, multivariate partial eta squared = .09, so Poster Quality can also be considered a measure of learning. Pairwise comparisons revealed a significant difference between Round 1 and each of the Rounds 2 and 3, but no significant difference between Rounds 2 and 3. It is possible that students found a good strategy on Round 2 and they did not change it substantially on Round 3. We identified the same pattern of results for Negative Feedback across the game: Round 1 = 1.53, Round 2 = 2.22, Round 3 = 2.04; Wilks' Lambda = .70, F(2, 87) = 19.01, $p < .001$, multivariate partial eta squared = .30. A reverse pattern was identified for Revision, students revising significantly less across the game rounds. Round 1 = .80, Round 2 = .66, Round 3 = .56; Wilks' Lambda = .81, F(2, 87) = 10.18, $p < .001$, multivariate partial eta squared = .19. We found that the shape of the revision decline trend was linear, not quadratic, suggesting a constant rate of individual decline in revision over time, slowing down from round 2 to round 3.

We conducted Spearman correlations on each poster round, because the variables involved in the analysis were also not normally distributed. On each of the second and third rounds, poster performance correlated with negative feedback choices as shown in Table **7** and, thus, inversely correlated with positive feedback. This indicates that the more the students choose negative feedback on rounds two and three, the better their posters are on those rounds.

Table 7. Spearman correlations between poster performance and choices on each game round

Measures (n = 89)		**Negative Feedback**	**Revision**
Round 1	**Poster Quality**	.20	.10
Round 2		**.27***	.08
Round 3		**.30****	**.33****

** $p < .01$ * $p < .05$

We did not find this relation on the first game round, perhaps because students were engaging in exploration on the introductory round and had not yet discovered a strategy for effective poster design. Importantly, looking at the first poster students created, before they received any feedback, there was no correlation between initial poster quality and the total amount of negative feedback chosen by students across the game (rho = -.08, p = .44). This lack of correlation on initial poster quality indicates that it is not just higher-performing students (i.e., performing well on their posters at the beginning of the game) who choose negative feedback and do well on their posters. In other words, choosing negative feedback is not serving as a proxy for incoming student diligence or ability.

1c. Do choices and time spent on feedback affect in-game performance?

We examined within-subject differences (i.e., comparing students to themselves) between choices and learning to further explore whether choosing negative feedback may help all students improve their performance. First, we investigated how students' choices of negative feedback were related to the time they spent reading the feedback. We examined the last two rounds of the game, when students presumably had settled on a learning strategy (i.e., choosing negative feedback and revising). The first round was fairly noisy because students were first introduced to the game and they were exploring its options and interface. We computed two measures, *Differential Feedback Dwell Time* and *Differential Negative Feedback*. To obtain the *Differential Feedback Dwell Time* measure, we subtracted the amount of time each student took to read the feedback on the second round from the amount of time that student took to read the feedback on the third round. Similarly, to obtain the *Differential Negative Feedback* measure, we subtracted the amount of negative feedback a student chose on the second round from the amount negative feedback that student chosen on the third round. That way, we could examine how a student's change in the amount of chosen negative feedback was related to that same student's feedback dwell time, essentially controlling for achievement level by comparing students to themselves at different time points. We conducted a Spearman correlation and found an association between *Differential Feedback Dwell Time* and *Differential Negative Feedback*, as shown in Table **8**. This indicates that a student who spends more time reading feedback from Round 2 to Round 3 also chooses more negative feedback from Round 2 to Round 3.

Next, we examined whether a closer read of the feedback also materialized into better poster performance. Therefore, we computed the *Differential Poster Quality* measure by subtracting a student's score on the first poster designed on Round 2 (before any revision) from that student's score on the last poster designed on Round 3 (after any revision). We found that *Differential Feedback Dwell Time* was positively associated with *Differential Poster Quality*, as shown in Table **8**. This indicates that a student who spends more time reading the feedback from Round 2 to Round 3 also shows a poster performance improvement from Round 2 to Round 3. This result implies that, indeed, paying attention to feedback drives learning and performance gains.

Table 8. Spearman correlations between dwell time, negative feedback, and performance from Round 2 to Round 3

Measures (n = 89)	**Differential Negative Feedback**	**Differential Poster Quality**
Differential Feedback Dwell Time	**.29****	**.24***

** $p < .01$ * $p < .05$

2. Examining the relation between students' choices and outside measures of their school achievement

The first set of analyses confirmed our hypothesis that in-game choices (negative feedback and revision) are associated with in-game performance (performance on the poster designs). Thus, students' behaviors while learning are as important as their performance outcomes and deserve much attention when developing assessments. The next question is whether assessing choices within the game reveals anything about performance outside of the game. This is equally important, because learning skills, such as seeking feedback and revising, may affect achievement in school. These analyses are organized into two main research questions:

a. Do choices to seek negative feedback and to revise correlate with in-school performance?
b. Does in-school performance mediate students' revision choices over time?

2a. Do choices to seek negative feedback and to revise correlate with in-school performance?

To investigate if in-game choices correlate with in-school learning outcomes, we used two types of measures: (1) school grades in Science and Mathematics collected during the academic year in which we conducted our study and (2) standardized test scores in English Language Arts and Mathematics collected two years before we conducted our study. The latter were employed because of a lack of standardized tests administered in California during the transition period from the STAR to the Common Core tests.

First, we employed students' Science and Mathematics grades. For these analyses, in addition to the students who did not complete all three posters, we excluded the students who did not provide consent for disclosing grades. Table **9** shows the relevant correlations. We found that the choice to seek negative feedback did not correlate with the Science Grade or with the Math Grade. However, Revision correlated with both in-school achievement measures (Science Grade and Math Grade). We did not find an association between poster performance (Poster Quality) and students' Science and Mathematics grades.

Table 9. Spearman correlations between negative feedback, revision, performance, and school grades

Measures	**Science Grade (n = 89)**	**Math Grade (n = 57)**
Negative Feedback	.15	.21
Revision	**.30****	**.30***
Poster Quality	.06	.04

** $p < .01$ * $p < .05$

Second, we employed students' STAR scores in ELA and Mathematics collected when the students were in grade 6. For these analyses, we excluded students who did not complete all three posters, as well as those who did not provide consent for STAR score disclosure. We found that negative feedback moderately correlated with the Math-CST score, but choices and in-game performance did not correlate with ELA-CST, as shown in Table **10**. We found no correlations between students' poster performance and students' grade 6 STAR scores in ELA and Mathematics. We tested the relations between all our measures and we found that they were linear.

Table 10. Spearman correlations between choices and STAR (Standardized Testing and Reporting) achievement scores in ELA and Mathematics collected two years prior to this study, when students were in grade 6

Measures (n = 75)	**ELA-CST**	**Math-CST**
Negative Feedback	.23	**.29***
Revision	.18	.19
Poster Quality	.19	.11

** $p < .01$ * $p < .05$

2b. Does in-school performance mediate students' revision choices over time?

In the previous analyses, we found that students' current Science and Math grades were highly correlated (*rho* = .75, *p* < .01), however, we used only the complete Science scores in our analyses, because a third of the students' Math grades were missing. We did not include the STAR scores in our further analyses because we did not find any correlations between Revision and students' STAR scores.

We conducted a new analysis to explore the relation between Revision and students' Science Grades. This analysis brings value to our initial correlational analyses, specifically because it can provide information regarding students' likelihood to revise at any point in the game and their revision trends, as well as reveal how other covariates (e.g., Science Grade) affect students' likelihood to revise over time. Hence, we can explore students' likelihood to revise at the beginning and throughout the game, as well as whether they revise at different rates across the game. We built three models incrementally, starting from the null model (Model 0) and adding fixed effects at each step. We found that Model 1 was a significantly better fit than Model 0, and that Model 2 was a significantly better fit than Model 1, as shown in Table **12**. The fit of the models was computed using the Laplace maximum likelihood approximation.

Model 0: We explored whether a GLMM method was needed to analyze the data. For this, we started with a null model (i.e., a model with a random intercept and no predictors) that partitions and quantifies the overall variance across *students*, to find out whether any of the variance of the model comes from individual differences, as shown in Table **11**. Thus, we computed the intraclass correlation (ICC), which constitutes the variance between individuals divided by the total variance (between and within individuals). The Level 1 variance for a Binomial distribution is estimated as $\pi^2/3$ (i.e., approximately 3.29; Heck, Thomas, & Tabata, 2013, p. 179). Thus, ICC = 1.786 / (1.786 + 3.29) = 41%, where the variance of Level 2 in Model 0 was 1.786. This value exceeds the 5% threshold, so we can assume statistically significant variability in intercepts between students, justifying the development of a multilevel model (Heck, Thomas, & Tabata, 2010, p. 25). Therefore, in Model 0, 41% of the total variance in initial revision comes from the students. The intercept constitutes the predicted log odds when all variables in the model are 0. The predicted probability (73%) based on the estimated intercept log odds coefficient in Model 0 (.994, *p* < .001; odds ratio = 2.7) can be interpreted as the percentage of students who revise across the entire game (time is not included in this model). In this case, the log odds intercept constitutes the grand mean log odds coefficient across all three game rounds. Thus, in Model 0, students are 2.7 more likely to revise than not to revise in Posterlet.

Model 1: We then explored all the combinations between fixed and random effects for this data set with the variables selected, but they did not improve the fit of the models presented in Table 12. We investigated other variables that could account for differences in Revision between students. We examined the relation between revision and in-school performance (Science Grade). Then, we added Science Grade as a covariate in Model 1 to explore how it affects students' likelihood to revise across the game and potentially explain some of the variance found in Model 0. We computed *ZScience*, the normalized Z-scores of Science Grade to facilitate result interpretation and we added it as a fixed effect to Model 0. We found that, in Model 1, 27% of the total variance in initial revision originates from the students. Results indicate a main effect for *ZScience*, indicating that students' Science grades are significantly related to students' likelihood to revise. The predicted probability (72%) based on the estimated intercept log odds coefficient in Model 1 (.932, $p < .001$; odds ratio = 2.54) can be interpreted as the percentage of students who revise across the entire game (*time* is not included in this model either). Thus, in Model 1, students are 2.54 more likely to revise than not to revise in Posterlet.

Model 2: As we mentioned before, we found that a model with a random intercept and a random slope did not improve the model fit, so we considered *time* a fixed effect. Model 2 shown in Table **11** is built by adding *time* and the interaction of *ZScience* with *time* as fixed effects to Model 1 to investigate whether the relation between students' Science Grades and revision varies between students over time. The *time* coefficient suggests that students' likelihood of revising decreases significantly (log odds = -.805, $p < .001$) over each time interval. It is possible that, as they improve their original designs, students do not feel the need to revise as often. It is also possible that, as the game progresses, the task is becoming tiresome. The predicted probabilities for all three levels of time can be obtained by first adding the log odds of the *time* and *intercept*. In Model 2, where we included the *time* fixed effect, the estimated intercept log odds coefficient was 1.860. For example, at time 0 (i.e., on Round 1), the log odds of revising are 1.860, with odds ratio $e^{1.860} = 6.42$ and an estimated probability of 86% (= 6.42/7.42), which constitutes the percentage of students who revised on Round 1 (i.e., students are 6.42 more likely to revise than not to revise at the beginning of the study). Similarly, the estimated probability for students who are likely to revise is 74% = 2.87/3.87 on Round 2 (the estimated log odds of revising are 1.055 = 1.860 – .805 and the corresponding odds ratio is 2.87), and 56% = 1.28/2.28 on Round 3 (the estimated log odds of revising are .25 = 1.860 - .805 - .805 and the corresponding odds ratio is 1.28). The correlation between the change (time slope) and the initial status (intercept) is negative (-0.764), indicating that students revise more initially but they tend to revise less over time. Results show no interaction between *ZScience* and *time*, suggesting that students' revising behaviors over time are *not* influenced by their science grades. Thus, lower-achieving students are not any less likely to revise than their higher-achieving peers. In Model 2, 35% of the variance in initial revising resides between students. This indicates that Model 2 can be refined further by exploring other covariates to better fit the data.

Table 11. Generalized linear mixed models built to predict the *Revision* dichotomous outcome

Model 0 (Null Model): Random effect – *intercept*	
$\eta_{ti} = \beta_{0i}$	(Level 1 Model)
$\beta_{0i} = \Upsilon_{00} + u_{0i}$	(Level 2 Model)
Model 1: Random effect – *intercept*. Fixed effect – *Science Grade*	
$\eta_{ti} = \beta_{0i} + \beta_{1i}$ (SCIENCE)	
$\beta_{0i} = \Upsilon_{00} + u_{0i}$	
$\beta_{1i} = \Upsilon_{10}$	
Model 2: Random effect - *intercept*. Fixed effects – *Science Grade*, *time, Science Grade*time*.	
$\eta_{ti} = \beta_{0i} + \beta_{1i}$ (SCIENCE) + β_{2i} TIME$_{ti}$ + β_{3i} (SCIENCE*TIME$_{ti}$)	
$\beta_{0i} = \Upsilon_{00} + u_{0i}$	
$\beta_{1i} = \Upsilon_{10}$	
$\beta_{2i} = \Upsilon_{20}$	
$\beta_{3i} = \Upsilon_{30}$	

Table 12. Revision (1=revised, 0=not revised) across time, GLMM fixed and random effect estimates with standard errors

Parameter	Model 0 (df=2)		Model 1 (df=3)		Model 2 (df=5)	
	β	**SE**	**β**	**SE**	**β**	**SE**
Intercept (Υ_{00})	**.994*****	.23	**.932*****	.21	**1.860*****	.36
SCIENCE (Υ_{10})			**.712*****	.20	**.910****	.31
TIME (Υ_{20})					**-.805*****	.22
SCIENCE*TIME (Υ_{50})					-.109	.19
Variance	Var(int)	AIC	Var(int)	AIC	Var(int)	AIC
Components	1.786	362.10	1.254	346.93	1.787	336.37

Var(int): Intercept Variance, AIC: Akaike Information Criterion

DISCUSSION, LIMITATIONS, FUTURE WORK, AND IMPLICATIONS

An assessment of students' choices provides a new approach for formative assessment and evaluating process skills that have been elusive to more traditional testing, but are of great interest to many educators. Traditionally, choice has been considered as simply a source of motivation for learning (Iyengar & Lepper, 1999). However, the results presented here, as well as in our previous work, make an argument for choice as an important factor that predicts learning (Cutumisu, Chin, & Schwartz, 2014). We examined the relation between students' in-game choices and their in-game and in-school learning performance. We described our assessment, including this manuscript's main contribution, the automatic feedback system, which enabled us to examine the relationships among feedback, revision, and learning performance. The study results show that the degree to which students choose negative feedback and revise correlates with multiple measures of learning. Moreover, we examined how different measures of learning influence students' probability to revise their work over time.

Choices and In-game Performance Outcomes. We investigated whether choices that students make in a short online game while designing posters could reveal anything about students' performance in the same environment, to solidify our use of choice as a central focus for new assessments of learning processes. Both choices to seek negative feedback and to revise correlated with in-game learning, though the regression analysis indicates that students' choice of negative feedback is a better predictor for the overall quality of the posters than revision. This result echoed our previous studies (e.g., Cutumisu, Blair, Chin & Schwarz, 2015; Cutumisu & Schwartz, 2014; Cutumisu & Schwartz, 2016) and it suggests that the choice to seek negative feedback may be more important for in-game performance than the choice to revise and that it should be pursued further. Finally, choosing negative feedback and choosing to revise one's poster were strongly correlated both across the game and particularly on the last game round, where students potentially found a stable strategy for successful poster design. Although we cannot determine whether negative feedback causes students to revise or whether students who revise are more inclined to choose negative feedback, the strength of the correlation between choices here and in all our previous research prompts more in-depth analyses in future studies.

Choices and Out-of-game Performance Outcomes. We investigated whether choices that students make in the game environment could reveal anything about students' performance outside our assessment environment. This analysis aimed to strengthen the validity of choice as a central construct in assessments that reach beyond a game medium. Both choices were positively associated with school measures of performance. As in our previous research (Cutumisu, Blair, Chin, & Schwartz, 2015), the correlations with outside measures were not as consistent across our choices. Moreover, we did not find an association between poster performance and students' current Science/Mathematics grades or their prior ELA/Math STAR scores. We expect that we can improve the precision and reliability of our in-game performance measures through multiple design iterations and refinement of our poster quality measure. The results indicate a need for more research to determine the factors and conditions that influence these relations. Specifically, we found that the more the children revised within the game-based assessment, the better their grades were in Science and Mathematics. The GLMM analysis independently confirmed this finding, Science Grade predicting Revision, consistent with the strength and direction of correlations reported in this study. This analysis suggests that students who achieve better science grades in school tend to revise more initially (e.g., perhaps because high-achieving students already know and use good learning strategies) but not over time (e.g., perhaps because their posters became better and they felt less need to revise). Other research supports the benefic impact of revision on performance. In a MOOC environment, researchers found that prompt feedback on drafts improves course outcomes (Kulkarni, Bernstein, & Klemmer, 2015). In more structured domains, such as mathematics, researchers found that immediate computer-generated feedback, combined with revision, improves grades (Heffernan et al., 2012). Other researchers found similar patterns of results between learning choices and learning outcome outside of the assessment environment. For example, Conati and Vanlehn (2000) found that when students are prompted to choose what kind of self-explanation to provide in a Physics tutor, they learn more from examples. Similarly, Aleven and Koedinger (2000) evaluated the effect of reason-giving in a Geometry tutor and found that students who were assisted by the tutor improved their test scores and transferred their reason-giving skills outside the cognitive tutoring environment. Finally, Roll et al. (2011) assessed students' learning choices regarding the use of online learning resources, such as requesting hints from the system or using the online glossary in a Geometry Cognitive Tutor intelligent tutoring system. They found that the online assessment of metacognition (i.e., students' choices) correlated with achievement outside the tutoring system, indicating that students who make better help-seeking choices learn more mathematics. Moreover, students' help-seeking skills improved and transferred to a new learning context (outside of the assessment environment) with lasting effects when the help-seeking tutor was no longer available.

We also found that the more the children chose negative feedback within the game, the better their scores were in their previous standardized state tests in Mathematics. Similarly, students' class grades in science were related to their revision choices. Importantly, these strategies can be taught to lower-achieving students to help them make better choices that can improve their learning outcomes. In a prior study, we showed that these two particular behaviors (choosing negative feedback and revising) can be taught and they are especially beneficial to lower-achieving students (Conlin et al., 2015).

Choices and Performance Over Time. We examined the factors that may influence students' revision choices over time. We found that students' Poster Quality and amount of Negative Feedback chosen increased significantly across game rounds, while their revising significantly lessened. The generalized linear mixed model analyses also revealed that students revised significantly less over time, which is a common result for growth models (Heck, Thomas, & Tabata, 2013, p. 186).

Alternative Explanations. This study was correlational, so it is possible that there are other variables that drive the correlations between feedback, revision, and learning performance. First, self-confidence and views of intelligence (fixed versus growth mindset) may influence students' choices between positive and negative feedback (Ehrlinger, Mitchum, & Dweck, 2016). For instance, the perception of a trait as fixed may lead to avoidance of negative feedback (Dunning, 1995). Second, it may be that students who are more diligent also create better posters, choose more negative feedback, and revise more. For example, compared to a growth mindset (an incremental theory of intelligence - the belief that intelligence can be developed over time), a fixed mindset (an entity theory of intelligence – the belief that intelligence is fixed) was found to be associated with decreased attention to corrective feedback or errors (Mangels et al., 2006). Our within-subjects analyses of the time students spent reading the feedback they chose revealed that students who chose more negative feedback from one round to another also performed better on the poster design task and spent more time reading the chosen feedback. This result indicates that all students who choose negative feedback read this feedback more closely, regardless of their particular mindset.

Second, another possible mechanism for choosing positive over negative feedback comes from construal level theory that describes the relation between psychological distance and the level of abstractness of people's thinking. Specifically, action construals (i.e., ways in which individuals perceive, understand, and interpret the world) emphasize that the level of abstraction at which people construe self-evaluative opportunities influences feedback-seeking preferences. For example, researchers found that low-level construals (i.e., more concrete thinking), as well as temporally proximal outcomes, elicit greater preferences for favorable (self-esteem protection) feedback. In contrast, high-level construals (i.e., more abstract thinking) and temporally distal outcomes lead to a preference for unfavorable (self-evaluation utility) feedback (Freitas, Salovey, & Liberman, 2001). Focusing on the *process* of self-evaluation may activate self-enhancement goals (low-level action construals), but focusing on the *purpose* of self-evaluation often may activate realistic-assessment goals (high-level action construals; Vallacher & Wegner, 1985, 1987; Liberman & Trope, 1998). Thus, some students may have focused on the process of choosing and reading the negative feedback that could potentially create discomfort, and hence chose positive feedback. Others focused on the utility of choosing negative feedback as a means to improve their posters and subsequently sell more tickets and, hence, chose negative feedback. We plan to interview students regarding their choices of negative feedback and revision in Posterlet to gain an insight into their decision-making processes.

Third, it was found that students' moods can play a role in feedback seeking (Trope & Neter, 1994); participants in a laboratory setting who were in positive moods sought more negative than positive feedback, while participants in bad or neutral moods showed either no preference or sought more positive than negative feedback. A future study could explore mood as a factor that could impact students' choices to seek negative feedback.

Finally, the importance of the domain could also play a role in people's feedback preferences and it may explain the variability in the correlations between choices and in-game versus out-of-game measures. For instance, Freitas, Salovey, and Liberman (2001) found that the domain (social versus achievement) and feasibility (the ease versus the desirability/utility of acquiring feedback) play a role in people's feedback preferences, which may affect differentially the relations of negative feedback with in-game performance and school grades. Perhaps some students considered the poster design game activity more academically relevant than other students and, thus, chose less negative feedback.

Limitations. Due to time constraints, we were unable to administer an outside post-test of graphic design principles as we have done in previous studies. Thus, the study lacks an offline pre-post gain measure in the same content domain (i.e., poster design). Instead, we employed students' achievement scores as measures that are independent from our assessment environment and domain content. Although we received students' grades in Science and Mathematics, grades are generally considered noisy measures, as they may be more susceptible to showing association through a moderator variable (e.g., general attempts). However, we also obtained students' prior ELA and Mathematics standardized test scores as more stable, supplementary in-school measures of students' academic achievement. Moreover, we conducted more powerful statistical analyses involving Science Grade to better understand the relation between students' achievement and choice to revise.

Implications. This research builds on our prior work to show that that automatic feedback generation systems can be successfully embedded even in an open-ended graphic design domain (e.g., digital poster design) to benefit learning performance. There are several possible implications of this work for both research and instructional purposes. First, the description of this intelligent feedback system enables other researchers to apply this framework to other open-ended domains. Second, the flexibility of such short assessments that are focused on specific choices (e.g., feedback seeking) enables the development and evaluation of a variety of instruction models. Third, negative feedback as a choice can be taught to students, now that we have evidence that it is associated with better learning outcomes. Moreover, these types of assessments may be used to evaluate students' willingness to seek constructive feedback, while simultaneously providing teachers with live reports automatically generated by the learning analytics already embedded in these systems. This would enable both teachers and students to use the feedback immediately. Most importantly, these assessments could offer teachers a window into how students interact with the feedback, so that teachers can adapt their instruction to best support students' learning.

CONCLUSIONS

We present a novel feedback system that evaluates the use of graphic design rules in an open-ended digital poster design domain and creates customized feedback within a choice-based game assessment environment. The game environment is designed for player agency with regards to learning choices and it captures students' willingness to seek negative feedback and to revise. The learning analytics embedded in the assessment environment quantified students' choice and poster performance, facilitating their comparison with external or in-school achievement measures.

We found that choice outcomes were directly related to traditional learning outcomes. Specifically, both choices correlated with performance in the game and with each other. The choice to revise correlated with students' current Science and Mathematics grades, while the choice to seek negative feedback correlated with their latest standardized Mathematics test recorded two years before this study. This work examines important learning strategies that instructors can provide to students and that students can employ to improve their learning, in the context of an assessment model centered around choice outcomes rather than solely on traditional learning outcomes. Importantly, the detailed description of the intelligent feedback system embedded into our assessment environment may aid other researchers in developing similar systems to track students' process behaviors and to compare the effects of different courses of instruction, regardless of a specific curriculum.

ACKNOWLEDGEMENTS

This research was generously supported by funding from the Gordon and Betty Moore foundation and the NSF (Grant # 1228831). We thank Neil Levine for his Posterlet graphics and for providing the 21 graphic design principles, which constituted the basis of our feedback system. We especially thank the students who participated in this study, as well as their teacher. We are grateful to the anonymous reviewers for their helpful suggestions that improved this manuscript.

REFERENCES

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In *International Conference on Intelligent Tutoring Systems*, 292-303. Springer Berlin Heidelberg.

Anseel, F., Beatty, A. S., Shen, W., Lievens, F., & Sackett, P. R. (2015). How are we doing after 30 years? A meta-analytic review of the antecedents and outcomes of feedback-seeking behavior. *Journal of Management, 41*(1), 318-348.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational Behavior and Human Performance 32*(3), 370-398.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-74.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. *Yearbook of the National Society for the Study of Education, Blackwell Publishing Ltd, 103*(2), 20-50.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61-100.

Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Second Edition. Hillsdale, NJ: Erlbaum.

Conati, C., & Vanlehn, K. (2000). Toward Computer-Based Support of Meta-Cognitive Skills: a Computational Framework to Coach Self-Explanation. *International Journal of Artificial Intelligence in Education, 11*, 389-415.

Conlin, L., 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. In *Proc. of the 122nd American Society for Engineering Education (ASEE'15)*, ms. 10 pages, June 14-17, Seattle, WA, USA. *Finalist in the Best of Design in Engineering Education (DEED) Division.*

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 245–252. New York, NY: ACM Press.

Core Team, R. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from: https://www.Rproject.org/.

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

Cutumisu, M., Chin, D. B., & Schwartz, D. L. (2014). A game-based assessment of students' choices to seek feedback and to revise. In *Proc. of the 11th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA '14)*, Porto, Portugal, October 25-27, pp. 17-24. Best Paper.

Cutumisu, M., & Schwartz, D. L. (2014). Choosing negative feedback improves learning for students of all ages: A game-based assessment of seeking negative feedback and revising. In *Proc. of the London International Conference in Education*, London, UK, November 10-12, pp. 171-176.

Cutumisu, M., & Schwartz, D. L. (2016). Choosing versus receiving feedback: The impact of feedback valence on learning in an assessment game. In *Proc. of the 9th International Conference on Educational Data Mining (EDM '16)*, June 29 - July 2, pp. 341-346, Raleigh, NC, USA.

Dunning, D. (1995). Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives. *Personality and Social Psychology Bulletin, 21*, 1297–1306.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), 256-273.

Ehrlinger, J., Mitchum, A. L., & Dweck, C. S. (2016). Understanding overconfidence: Theories of intelligence, preferential attention, and distorted self-assessment. *Journal of Experimental Social Psychology*, *63*, 94-100.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review, 100*(3), 363-406.

Freitas, A. L., Salovey, P., & Liberman, N. (2001). Abstract and concrete self-evaluative goals. *Journal of Personality and Social Psychology*, *80*(3), 410.

Hattie, J. (1999). Influences on student learning. *Inaugural Lecture*: Professor of Education, University of Auckland, New Zealand, August 2.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Heck, R. H., Thomas, S., & Tabata, L. N. (2010). *Multilevel and longitudinal modeling with IBM SPSS ebook*. Routledge Academic.

Heck, R. H., Thomas, S., & Tabata, L. N. (2013). *Multilevel modeling of categorical outcomes using IBM SPSS*. Routledge Academic.

Heffernan, N., Heffernan, C., Dietz, K., Soffer, D., Pellegrino, J. W. Goldman, S.R., & Dailey, M. (2012). Improving Mathematical Learning Outcomes Through Automatic Reassessment and Relearning. *AERA*.

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: a cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology, 76*(3), 349-366.

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), pp. 978-992.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254-284.

Kluger, A. N., & DeNisi, A. (1998). Feedback Interventions: Toward the Understanding of a Double-Edged Sword. *Current Directions in Psychological Science*, *7*(3), 67–72.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30-43.

Kulik, J. A., & Kulik, C.-L. C. (1987). Timing of Feedback and Verbal Learning. *Review of Educational Research 58*(1), 79–97.

Kulkarni, C. E., Bernstein, M. S., & Klemmer, S. R. (2015). PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 75-84. ACM.

Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of personality and social psychology*, *75*(1), 5-18.

Luminance Algorithm. (2016). http://www.w3.org/TR/2008/REC-WCAG20-20081211/#relativeluminancedef.

Mangels, J. A., Butterfield, B., Lamb, J., Good, C., & Dweck, C. S. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Social Cognitive and Affective Neuroscience, 1*(2), 7586.

Mitrovic, A., & Ohlsson, S. (2016). Implementing CBM: SQL-Tutor After Fifteen Years. *International Journal of Artificial Intelligence in Education*, *26*(1), 150-159.

Nussbaum, A. D., & Dweck, C. S. (2008). Defensiveness vs. remediation: Self-theories and modes of self-esteem maintenance. *Personality and Social Psychology Bulletin, 34*, 127–134.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Roll, I., Aleven, V., McLaren, B., & Koedinger, K.R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 267–280.

Roll, I., Baker, R. S., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. Journal of the Learning Sciences, 23, 537–560.

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

Schwartz, D. L., Tsang, J. M., & Blair, K. P. (2016). *The ABCs of How We Learn: 26 Scientifically Proven Approaches, How They Work, and When to Use Them*. W. W. Norton & Company.

Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology, 39*, 212–222.

Trope, Y., & Neter, E. (1994). Reconciling competing motives in self-evaluation: The role of self-control in feedback seeking. *Journal of Personality and Social Psychology, 66*, 646–657.

Vallacher, R. R., & Wegner, D. M. (1985). *A theory of action identification*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological review*, *94*(1), 3-15.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3–17.