Posterlet: A Game-Based Assessment of Children's Choices to Seek Feedback and to Revise

Maria Cutumisu, Kristen P. Blair, Doris B. Chin, Daniel L. Schwartz

Graduate School of Education, Stanford University, USA

cutumisu@ualberta.ca

ABSTRACT: We introduce one instance of a game-based assessment designed to measure students' self-regulated learning choices. We describe our overarching measurement strategy and we present *Posterlet*, an assessment game in which students design posters and learn graphic design principles from feedback. We designed *Posterlet* to assess children's choices to seek informative negative feedback and to revise their work. Middle-school students from New York and Illinois played *Posterlet* and then took a post-test, for an overall average of 17 minutes of interaction time. Results showed that the frequency of choosing negative feedback and revision correlated with learning graphic design principles from the game. Seeking negative feedback, but less so revision, further correlated with standardized achievement tests of reading and mathematics. Our research presents a first-of-kind behavioural measure of students' feedback and revision choices and their relations to learning. Within the design context of creating posters, we found correlational evidence that seeking negative feedback and revising are good behaviours for self-regulated learning, and we devised a way to measure these behaviours. This sets the stage for developing and evaluating models of self-regulated learning instruction that help students choose to seek feedback effectively and revise accordingly.

KEYWORDS: self-regulated learning, preparation for future learning, choice, assessment, game, learning, feedback, revision

1 INTRODUCTION

This article describes our initial efforts to develop and empirically validate game-based assessments of self-regulated learning (SRL). Here, we capitalize on the interactive, data-logging possibilities of learning technologies to capture process data about students' selection of positive or negative informative feedback. We demonstrate that the choice to seek constructive critical feedback correlates with greater learning within our specific assessment environment. It also correlates with multiple standardized measures of academic achievement. This work begins to realize Butler and Winne's proposal that "research on feedback and research on SRL should be tightly coupled" (1995, p. 248).

We begin by discussing general issues of assessing SRL and our general solution of choice-based assessments. We then briefly review the literature on feedback to identify the empirical gap that our assessment addresses. Next, we describe in some detail the focal assessment, called *Posterlet*, and we present results of a correlational study using a sample of convenience from New York and Illinois. Finally, we summarize the findings, their limitations, and possibilities for future work.

1.1 Assessing Self-Regulated Learning

A major goal of formal and informal education is to prepare students to be autonomous learners who have the "will and skill" to learn without the strict guidance of a parent, teacher, or computer (Pintrich & De Groot, 1990, p. 38). The goal of fostering independent learners falls within the research traditions of SRL and self-directed learning. For example, Garrison states, "self-directed learning is also a necessity if students are to learn how to learn and become continuous learners" (1997, p. 29).

A frequently acknowledged impediment to developing both theories and practices of SRL is the lack of precise behavioural measures (e.g., Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007; Howard-Rose & Winne, 1993). The authors of the concluding chapter of the 2000 *Handbook of Self-Regulation* state, "current research relies heavily on self-report measures. Thus, more observational and performance measures relevant to self-regulation processes and outcomes are urgently needed. Because there is a fundamental problem with using self-reports and survey methods to demonstrate dynamic processes, we sorely need better ways to operationalize the self-regulation construct" (Zeidner, Boekaerts, & Pintrich, 2000, p. 757).

When we consider existing behavioural assessments, the situation is not ideal. Most tests are poorly suited to evaluating SRL effects. They measure whether students have mastered the content, but they do not capture the SRL behaviours that may or may not have produced the measured learning. They also miss the central concern of SRL instruction, which we view as preparing students for future learning (Bransford & Schwartz, 1999). An ideal assessment would present an environment where students have something to learn, and then evaluate the processes students undertake to complete that learning.

Existing preparation for future learning (PFL) assessments, however, have predominantly focused on knowledge rather than process outcomes. In the typical PFL study, researchers compare which of two forms of instruction better prepares students to learn from an expository resource, such as a lecture, a reading, or a worked example (e.g., Schwartz & Martin, 2004). This determines which instruction prepared students to learn from new explanations, but not the processes students employed to learn. The maturation of computer-based learning environments, as well as rapidly evolving data-analysis techniques, have opened new possibilities for the collection and analysis of student process data (e.g., Stevens & Thadani, 2007). To expand PFL assessments to measure SRL behaviours directly, we are developing simple games in which students are presented with challenges to solve, and we measure students' SRL choices.

1.2 Choice-Based Assessments

Independent learners need to make choices with the imperfect information they have at their disposal. For example, students need to choose what and how to learn. In many educational discussions, student choice is seen as a way to increase motivation and learning during instruction (Iyengar & Lepper, 1999). Here, we take choice as an outcome of instruction. Measuring choice as an outcome may be an

important proxy for students' learning in the future, when they must learn on their own.

Our measurement of student choice follows an intent-to-treat logic. Intent-to-treat studies test both whether a particular course of action yields a desirable result and whether people will take that course of action. For example, intent-to-treat studies of a new medication not only ask whether the medication improves patient outcomes, but also whether patients will choose to take that medicine. Similarly, given the practical goals of SRL instruction, the desired outcome is the prescription of useful strategies that students also choose to use.

We design our assessment game environments to follow three measurement principles:

1) *Preparation for Future Learning*. As mentioned above, assessments need to include opportunities to learn, so we can measure whether students employ SRL to learn the available content.
2) *Choice-Based*. Choices about learning need to be free choices, not necessarily right or wrong choices. In other words, the assessment environment must be pedagogically agnostic (i.e., it cannot tilt children towards one choice or another) and allow children to advance in the game regardless of their learning choices.
3) *Typical Learning Performance*. Students often exert maximal performance for high-stakes tests, which may not reflect the choices they would make for everyday learning (Sackett, Zedeck, & Fogli, 1988; Klehe & Anderson, 2007). Therefore, we create assessment environments that do not feel like tests, under the assumption that typical performance is a better proxy for independent learning behaviours, where there is no test to drive performance.

Our game-based assessments share many similarities to other learning technologies that collect and analyze process data. However, there are also some notable practical differences. One primary difference is that we are designing stand-alone assessments. The assessments are not embedded within a specific curriculum, and therefore, they can be used to compare the effects of different courses of instruction. Many computer-based learning environments, such as intelligent tutoring systems (e.g., Koedinger, Anderson, Hadley, & Mark, 1997), measure learning as students work through lessons within the system. These systems can evaluate student knowledge growth over time, aiming to determine the best sequence of problems or tutor moves that will maximize learning from the tutor (Chi, VanLehn, & Litman, 2010). Recently, this work has further begun to consider affect (e.g., boredom and engaged concentration), gaming the system, and off-task behaviours (Fancsali, 2014; Baker et al., 2008; Baker, Corbett, & Koedinger, 2004; Baker, Corbett, Roll, & Koedinger, 2008; Roll, Baker, Aleven, & Koedinger, 2014). The primary purpose of the measurements is to determine ways to help students learn more from the system. This is of high importance, but it constitutes a different goal than examining whether various forms of SRL instruction prepare students to learn in new contexts.

For instance, in the context of an intelligent tutoring system (ITS), Roll and colleagues (Roll, Aleven, McLaren, & Koedinger, 2011a, 2011b) found that automated tutoring on help-seeking and self-assessment behaviours helped students transfer these behaviours to another lesson within an ITS

environment. However, there were no available external assessments to determine whether these SRL behaviours would transfer beyond an ITS. Therefore, we hope to help researchers determine whether students transfer SRL behaviours from initial instruction to a new setting of learning by creating a suite of targeted choice-based assessments.

A second difference is that choice-based assessments have *a priori* measurement goals close to the surface rather than inferred through extensive data manipulation. Data-mining techniques have been employed to examine students' learning choices in games, often focusing on predicting future student moves or likelihood of success within the system (Peacock et al., 2013; Lee, Liu, & Popović, 2014; Snow, Allen, Russell, & McNamara, 2014). However, a communicative challenge of data-mining techniques is that the claim connecting behaviours to constructs often requires a complex chain of data transformations, difficult for stakeholders to interpret (e.g., a clustering algorithm that creates an abstract centroid in a multi-featured space). Hence, we have tried to close the distance between the raw behaviours and their assessment interpretations (e.g., by simply counting a student's frequency of choosing negative versus positive feedback). Ideally, making the construct more transparent will aid in educator decision making.

A third difference is that we try to keep the assessments relatively short. In one assessment study, Shute, Ventura, Bauer, and Zapata-Rivera (2009) had an *a priori* assessment goal of measuring student persistence using the physics learning game, Newton's Playground (Ventura & Shute, 2013). They demonstrated convergent validity by correlating their assessment results against existing measures of persistence. However, the assessment depended on several hours of student data, in part because the goal was to adapt the game dynamically to address individual student persistence needs. Given the many demands on classroom time, a multi-hour assessment designed to measure one aspect of SRL is unlikely to scale well. Thus, we have opted for assessments of 10–15 minutes. This increases the flexibility of deployment, although we do pay the price of having less information about any given student.

Ultimately, we aim to build a suite of choice-based assessments to examine the transfer of SRL behaviours, including the choices to engage in critical thinking (Chi, Schwartz, Chin, & Blair, 2014), to read to learn (Chin, Blair, & Schwartz, 2015), and to persist after failure (Chase, Chin, Oppezzo, & Schwartz, 2009). Ideally, these and more choice-based assessments can indicate whether specific educational experiences foster independent learners, the main desiderate of SRL research.

Here, we describe research on a game-based assessment called *Posterlet*. We designed *Posterlet* to measure two behaviours purported to be important for learning: students' *choices to seek negative feedback* and *to revise* (Cutumisu, Blair, Schwartz, & Chin, 2015). We posed three primary research questions to explore the relation between learning choices and learning outcomes:

- Do choices to seek negative versus positive feedback, and choices to revise, correlate with learning from the game?

- Do feedback and revision choices within the game correlate with broader learning outcomes outside the game (i.e., standardized achievement outcomes)?
- Do patterns of choice differ by school, suggesting that SRL choices can be influenced by experience?

2 LEARNING FROM FEEDBACK

The literature on feedback is voluminous and permits many different frameworks (e.g., based on types of feedback [Mory, 2004] or based on functions of feedback [Butler & Winne, 1995]). Different research traditions use different criteria for what counts as feedback, because feedback may refer to any signal relevant to one's thoughts, behaviours, or products. In behaviourist theory, feedback strictly refers to positive or negative reinforcement (e.g., reward and punishment). There was even a period when researchers debated whether "knowledge of results" comprised feedback (i.e., information indicating how to adjust behaviour [Schmidt, Young, Swinnen, & Shapiro, 1989]). Conversely, in control theory, feedback only refers to information that indicates the degree of discrepancy between a current and a desired state (Powers, 1978). In the SRL tradition, further distinctions arise when feedback can refer to internally generated signals, as might be gained from self-monitoring during self-explanation activities (Butler & Winne, 1995), or to external feedback generated by the environment or by another person (Zimmerman, 1990; Okita & Schwartz, 2013), as well as whether the feedback is task- or person-directed (Hattie & Timperley, 2007).

Blurring across the many possible distinctions, Kluger and DeNisi (1998) performed a useful meta-analysis comprising 131 studies. Feedback improved performance by 0.4 standard deviations on average compared to no feedback controls. Of special importance to our investigation, one-third of the studies actually found feedback to be worse than no feedback at all. The prevailing explanation for the benefits and drawbacks of feedback emphasize negative feedback. The authors found that negative feedback was 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 change. Yet, despite the signal for a need to change and learn, negative feedback runs the risk of triggering ego threat issues that lead people to shut down rather than heed the feedback (Hattie & Timperley, 2007). This suggests that students' attitudes towards negative feedback could have large implications for learning. Zimmerman highlighted this possibility among his list of three critical features of students' self-regulated learning: "their use of self-regulated learning strategies, their *responsiveness to self-oriented feedback about learning effectiveness*, and their interdependent motivational processes" (1990, p. 6; italics added for emphasis). Hence, we thought it important to develop an assessment that measured student behaviours and learning with respect to positive and negative feedback. This way, we can closely examine the relation of negative and positive feedback to learning, and we can determine whether there are important individual SRL differences.

It is noteworthy that the feedback literature was launched from the behaviourist tradition (Thorndike, 1927). Nearly all learning studies of external feedback use supervised feedback, where the teacher, experimenter, or computer decides when and how to deliver feedback. Learners do not have independent control — the feedback arrives without choice. However, in many situations, people need

53

to choose to seek feedback. For example, when writing a manuscript, authors may (or may not) ask their peers to read it before submission. They may choose peers who they know provide generally supportive comments or they may choose peers who are more likely to provide strong critique.

There is scant literature on people's feedback-seeking behaviours. For example, to our knowledge, research on the effects of immediate versus delayed feedback and their impact on learning, self-monitoring, and motivation has not included an opportunity for students to choose between immediate or delayed feedback (e.g., Schooler & Anderson, 1990; Lepper, Woolverton, Mumme, & Gurtner, 1993). Work on self-monitoring (e.g., self-explanation [Chi, 2000]) emphasizes individuals' abilities and willingness to detect inconsistences in their own thoughts, not their choice to receive feedback. Closer to the question is the work by Roll and colleagues (Roll, Baker, Aleven, & Koedinger, 2014) on help-seeking behaviours in ITSs, examining students' tendencies to ask for the right answer rather than to fix their understanding based on partial hints. Their research examines how learner choices vary as a function of the informative value of the hints, rather than of the feedback valence. Corbett and Anderson (2001), using a cognitive tutor context for Lisp programming, examined the timing of tutorial advice. They investigated whether various combinations of control over error correction and feedback/help (immediate/on-demand, as well as right/wrong, informative, or none) influenced learning outcomes. However, they only found that the various tutor conditions, when combined, improved performance compared to a system that just provided right/wrong information after the submission of a complete program. An empirical question remains whether help seeking and feedback seeking belonged within the same theoretical SRL construct. Creating an assessment of the choice to seek positive or negative feedback could help refine and differentiate how these behaviours fit within theories of SRL.

The most directly relevant prior work on choices to seek positive and negative feedback comes from consumer research. Finkelstein and Fishbach (2012) compared novices and experts (true domain experts and self-perceived experts). They found that novices sought more positive feedback, whereas experts sought more negative feedback. The demonstration that greater (perceived) expertise correlates with greater selection of negative feedback points to possible causal mechanisms (e.g., self-efficacy). Unfortunately, they did not measure learning outcomes, so the relation to SRL remains undetermined.

In sum, there are reasons to believe that behaviours and attitudes regarding negative feedback influence learning, but there is no evidence to date whether independent choices about seeking feedback are important for learning. Thus, we decided to investigate this question by creating *Posterlet*.

3   THE ASSESSMENT ENVIRONMENT: *POSTERLET*

Posterlet is a game-based assessment that enables students to design posters for their school's Fun Fair. The environment offers students the choice between negative and positive feedback to help them learn about graphic design principles. It also measures whether students choose to revise their work after feedback. Behaviours after feedback, such as revising and help seeking, can be important aspects of learning because they enable students to practice the correct skill. On the other hand, revision may be

primarily useful for improving the final product, but not for learning.

Figure 1 summarizes a cycle of *Posterlet*. On each of the three cycles of the game, students start the round by choosing an activity booth: basketball toss, science of bubbles, estimation station, pie-baking, or electricity (Step 1).
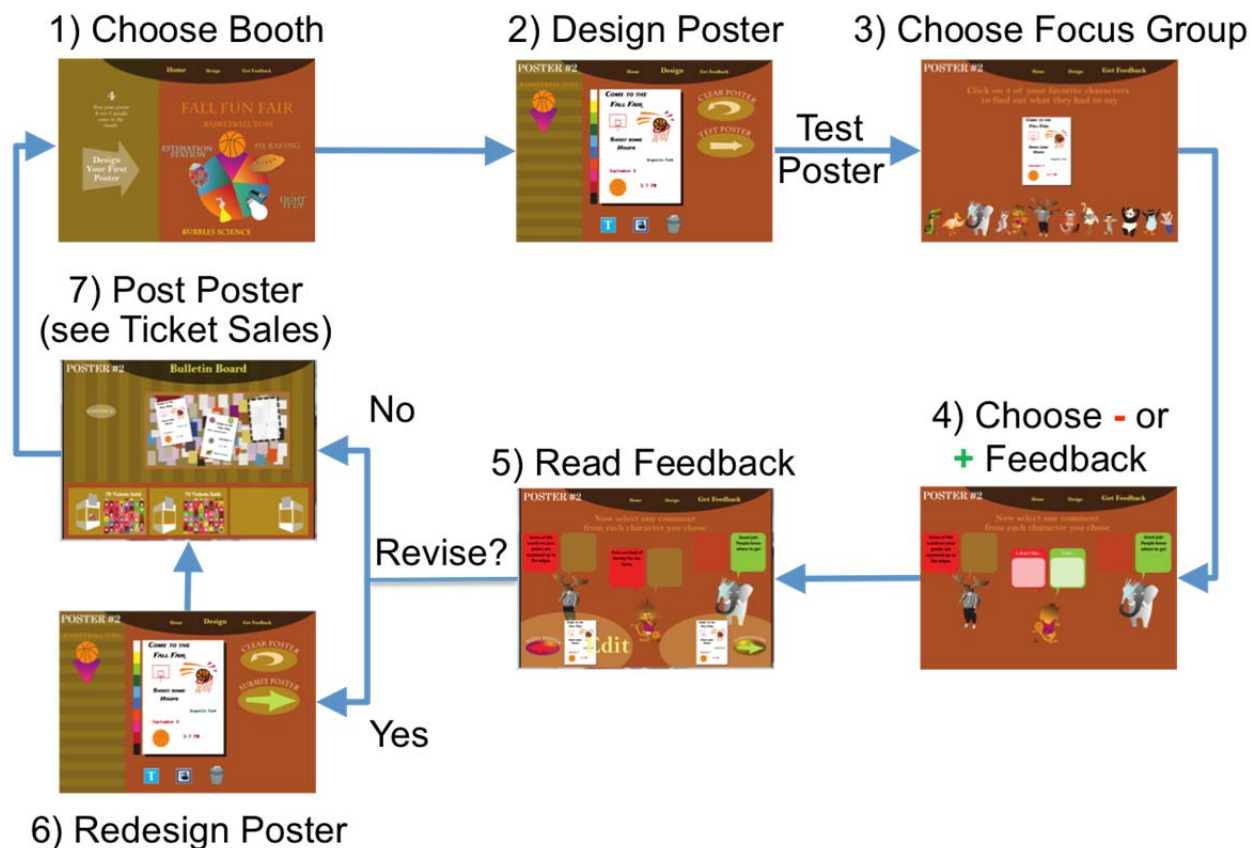

Figure 1. *Posterlet*: a game to assess students' learning choices in a design context. This diagram shows the state of the poster design cycle for the second poster.

In Step 2, students design a poster for their chosen booth. Students can select predefined phrases from a text palette and images from an image palette. In addition to placing text and images wherever they choose on the poster's canvas, students may customize the appearance of text (size, font, alignment, and colour), resize the images, and change the poster's background colour.

After designing their poster, students select three focus-group members from a cast of twelve animal characters (Step 3). Figure 2 shows how students choose either positive or negative feedback from each member of their focus group (Step 4).

Figure 2. Students may choose either positive ("I like…") or negative feedback ("I don't like…") from each of three characters they selected from the focus group. Here, the student has chosen one positive and one negative feedback, and has not yet chosen feedback from the third character.

Here, feedback is delivered by computer characters so, in a mild sense, it constitutes social feedback. Both the negative and the positive feedback refer to the execution of graphic design principles, not to artistic flair (see Table 1). Positive and negative feedback are equally informative and non-directive. In *Posterlet*, negative feedback is non-punishing. Moreover, we are measuring the effects of the feedback's valence (positive versus negative), not of the feedback's informational content (informative versus uninformative), on learning.

Table 1. Examples of positive and negative feedback phrases for each type of feedback (informative or uninformative) and each category (information, readability, or space use)

Type	Category	Positive	Negative
Informative	**Information**	It's nice that the poster says how much the booth costs.	You didn't say how much the booth was.
	Readability	Your poster has big letters. Really easy to read.	People need to be able to read it. Some of your words are too small.
	Space Use	Great job on getting your pictures away from the edge.	Your picture is squished up next to the edge.
Uninformative	**Generic**	Yay! Fairs are fun!	Hmm, I don't really like fairs.

After reading the feedback (Step 5), students must choose whether to revise their poster (Step 6) or submit it (Step 7). Once students submit their poster, they can see how many people buy tickets for their booth. The game tracks and displays a poster quality score as the number of tickets sold at each booth, potentially providing extra engagement and motivation for students to improve their scores. Students complete this poster design cycle for three booths. The students do not have an opportunity to seek feedback on a revised poster. Therefore, in total, they have nine decisions for positive or negative feedback (three per poster), and three decisions for revision (one per poster).

While students may learn simply through the process of designing and reflecting on their posters, the feedback provides specific information for learning 21 graphic design principles gathered from a consulting professional graphic artist (Figure 3).

INFORMATION	READABILITY	SPACE USE
location absent	text cut off by bounding box	space used by graphics out of range
date absent	text size small	top half empty
time absent	text style unreadable	bottom half empty
ticket price absent	text & graphics overlap	graphics touching edge
booth description absent	text and other text overlap	text flush with edge
graphics irrelevant	graphics size small	text flush with other text
text absent	text contrast low	-
images absent	-	-

Figure 3. The three categories of the 21 graphic design principles used by the feedback system to generate feedback: information, readability, and space use

These principles correspond to the three categories of informative feedback: pertinent information, readability, and space use. A graphical analysis system embedded in the game evaluates in real-time whether each principle is satisfied for any given poster. For instance, to detect a principle about colour contrast, we implemented an algorithm that computed the contrast and relative luminance[1] between the text colour and the background colour. The system generates poster-specific feedback, employing a prioritization scheme to select which principle to emphasize in the feedback. The scheme ensures a balanced coverage of the three feedback categories and does not repeat prior feedback. In situations when all eligible design principles have been exhausted, the system provides an uninformative generic phrase (see Table 1).

---

[1] http://www.w3.org/TR/2008/REC-WCAG20-20081211/#relativeluminancedef

4   THE EMPIRICAL INVESTIGATION

A unique feature of choice-based assessments, compared to knowledge-based assessments, is the burden to validate choices (i.e., to establish whether some SRL choices are in fact better than others for learning). Unlike assessments of knowledge, where the correct answer for 2+2 is 4 and not 5, choices face a special challenge. The reason becomes apparent upon a close read of Pintrich and De Groot's seminal SRL article in which students self-reported their uses of cognitive strategies. One of the survey items asked, "When I study for a test I practice saying the important facts over and over to myself" (1990, p. 40). This is a poor cognitive strategy for learning, yet the researchers considered it an indicator of effective cognitive strategy use. To avoid this type of mistake, it is important to have evidence that a proposed SRL strategy is indeed good for learning. In the case of choosing negative feedback, there is no relevant prior literature, so we need to show that it is better for learning, at least in this instance. This is the major goal of the current research.

Several hundred middle-school students completed *Posterlet*. We logged their choices regarding feedback and revision. We correlated these choices with their uses of graphic design principles in their posters and their abilities to judge a poster on a subsequent post-test. We also correlated *Posterlet* behaviours with standardized achievement scores to determine if the choices *within Posterlet* have a relation to academic outcomes that occur *outside Posterlet* in very different contexts.

4.1   Method

4.1.1   Participants and Procedures

Data were collected in the spring of 2013 from students at two public middle schools in urban settings; one school was in New York City (NYC), the second in Chicago. These two schools implement a game-based pedagogy designed to instil a variety of 21st-century skills and attitudes. *Posterlet* was one of several assessments administered by external school evaluators. We do not have access to the results of the other assessments. In New York City, we collected game logs from 278 children in Grades 6 to 9 (mean age=12.1 years, SD=1.0). Of these children, 231 further completed the post-test of graphic design principles. We received standardized test scores for 119 students. From those students who indicated their ethnicity (n=137), 17% were African-American, 47% Caucasian, 29% Hispanic, and 7% identified as "other." In Chicago, we collected the logs of 203 children in Grades 6 to 8 (mean age=12.2 years, SD=0.7). Of these children, 194 completed the post-test. We received standardized test scores for 65 students. For those who self-identified, the ethnicity breakdown was 63% African-American, 8% Caucasian, 27% Hispanic, and 1% "other." Finally, we also collected log files for an additional 36 children from these two schools, without indication of the particular school to which they belonged. Standardized scores were unavailable for these 36 children and only 29 of them completed the post-test.

To maximize the available data for each analysis, we used the subset of students who had complete data for each specific research question (rather than only analyzing students who had complete data on all measures). For example, to analyze the relations among feedback, revision, and poster performance, we

58

began with the full sample. However, to analyze the relations among feedback, revision, and the post-test, we used the subset of students who completed *Posterlet* and the post-test.

In their regular classrooms, participants individually played the *Posterlet* game by designing posters of their choosing for three of the five Fun Fair booths. They then took the subsequent post-test. We collected data online automatically from both the game and the post-test. NYC students spent an average of M=12.2 minutes (SD=5.9) on *Posterlet* and an average of M=3.2 minutes (SD=2.7) on the post-test. Chicago students spent an average of M=14.9 minutes (SD=6.2) on *Posterlet* and an average of M=4.0 minutes (SD=2.3) on the post-test. Not all students completed the posters or the post-test due to time considerations. We were not present during *Posterlet* administration, so we relied on game duration to remove students who were ±2 SD from the mean. Our logic was that we should be conservative and eliminate any student who may not have followed the proctors' rules. This pruned the data pool by 9.1%: *Posterlet* corpus (n=272 for NYC, n=172 for Chicago) and post-test corpus (n=226 for NYC, n=163 for Chicago). Our results persisted even when we included the outliers in the analyses.

4.1.2   Measures

We employed measures of students' learning-relevant choices, learning of the graphic design principles, and academic achievement.

4.1.2.1   Measures of Learning-Relevant Choices

We focus on two learning choices: *negative feedback* and *revision*. Negative feedback refers to the number of times a student chose the negative feedback option ("I don't like…"). Students had nine total feedback choices, so the number of positive feedback choices ("I like…") is necessarily the inverse. Revision tallies the number of times students chose to revise their posters out of a maximum of three (one per poster).

4.1.2.2   Measures of Learning the Graphic Design Principles

We collected two indicators of how well students learned the graphic design principles: *Poster Quality* and *Post-test*. Poster Quality is a performance measure computed from the quality of the student posters and was determined by the same graphical analysis system that generated feedback. For each poster, the system scored each of the 21 graphic design principles by providing one point for each good feature (i.e., a feature always used correctly on that poster), subtracting one point for each poor feature (i.e., a feature used incorrectly on that poster), and providing zero points for features that were inapplicable or not present. Each poster score is the sum of the scores for each of the 21 graphic design principles and ranges from -21 to 21. The scores for the final poster on each round (revised poster, or initial poster if no revision) were combined to create an overall poster quality score.
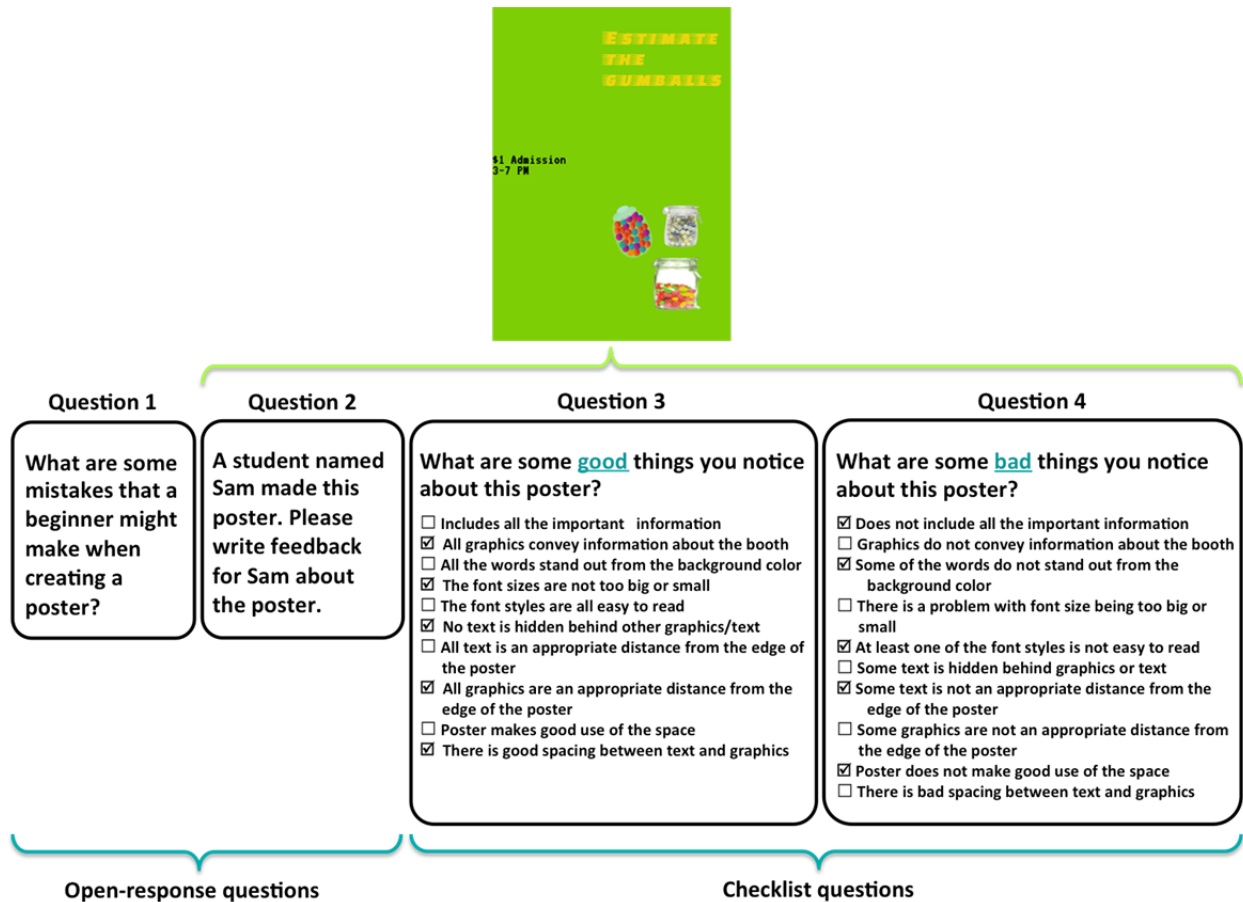
Figure 4. The four post-test questions: the items checked are the correct answers for Question 3 and Question 4, respectively

The post-test comprised four questions, as shown in Figure 4. The first open-response question asked students to list mistakes a beginner might make when creating a poster. The second open-response question provided a target poster and asked students to write feedback. We scored students' responses by counting the number of graphic design principles implied in each response (max=21). Open responses were coded individually by two evaluators with an inter-rater reliability score of r>.8. Two checklist items referred to the same target poster. In one item, students checked all the good things about the poster. In the other item, students checked all the bad things about the poster. Students received one point for correct answers that they did not otherwise contradict across the "what is good" and "what is bad" questions. The answers to these two questions are complementary, five being correct and five being incorrect for each question. We computed a total post-test score by adding the normalized scores (Z-scores) for each of the four questions.

4.1.2.3   Measures of Academic Achievement

For a subset of the students, we received Mathematics and English Language Arts achievement scores based on their respective state standardized tests: New York State Testing Program (NYSTP) and Illinois

Standards Achievement Test (ISAT). The fact that the students from the two schools had different standardized achievement tests makes direct comparison of achievement data across schools impossible. Nevertheless, if we find that the correlations between the choice variables and the standardized achievement scores are similar across schools and tests, this would indicate a potentially stable relation between choice behaviours within *Posterlet* and multiple broad measures of school learning.

4.2 Results

Across the game, students improved their poster quality, which we take as evidence that they learned; round 1=8.4 (SD=5.6), round 2=9.4 (SD=5.2), round 3=9.6 (SD=5.3); F(2,471)=14.9, *p*<.001. The question then is whether their choices regarding negative feedback and revision correlated with learning from the game and, more far reaching, with learning outcomes in school.

We describe the results in four sections, which we preview here. The first section considers whether in-game choices correlate with learning graphic design principles. It shows that, in the context of *Posterlet*, the choice to seek more negative feedback and the choice to revise independently correlate with student learning from the game. This is a first-of-kind demonstration, which warrants future testing to determine if this relation holds for other contexts and tasks, as well as for the causal variables driving the correlations. The second section considers the effect of choosing negative feedback on time spent reading feedback. It demonstrates that the more negative feedback students choose for a given poster, the longer they dwell on the feedback. This may further help explain how negative feedback, at least when self-selected, contributes to greater learning. The third section evaluates whether the *Posterlet* results correlate with academic achievement more broadly. We could not collect *in situ* measures of feedback-relevant behaviours in school and correlate those measures with *Posterlet*. Nevertheless, the stable correlations between negative feedback in *Posterlet* and standardized achievement provide promissory evidence that seeking negative feedback in *Posterlet* may have a meaningful association with how well children are doing in school. Finally, the fourth section demonstrates that *Posterlet* can detect differences between the two school samples in their choice of negative feedback. While we do not know the causes of these differences, it does indicate that *Posterlet* may be deployable as a way to evaluate the effects of different courses of instruction (e.g., instruction that teaches SRL behaviours).

4.2.1 In-Game Choices Correlate with Learning Graphic Design Principles

Table 2 shows the zero-order correlations between frequency of negative feedback selection, revision, overall poster quality, and post-test score. The more negative feedback students chose, the better they performed on the post-test and on the overall poster quality. Given that the score for positive feedback is the inverse of negative feedback, this result indicates that seeking positive feedback is negatively correlated with learning to the same degree. It is important to note that these results do not mean that students who prefer positive feedback did not learn from that feedback. They just learned less than their counterparts who sought more negative feedback.

The correlation between revision and overall poster quality is to be expected, given that the overall poster quality score comprised revised (and improved) posters for those students who chose to revise. More importantly, the choice to revise positively correlated with post-test scores, which did not depend on the quality of the posters that students produced.

Table 2. Correlations between negative feedback, revision, and internal learning outcomes

Measures	Negative Feedback	Revision	Poster Quality	Post-test
		n=473	n=473	n=414
Negative Feedback	--	.47***	.28***	.23***
Revision		--	.34***	.24***
Poster Quality			--	.39***

Note: *** *p*<.001

Is student prior knowledge responsible for these effects? Perhaps students who knew more about designing posters sought more negative feedback, revised more, and performed better on the post-test. To investigate this possibility, we treated the students' first poster (prior to any feedback or revision) as a pre-test. When controlling for differences in pre-test scores, all the correlations remain significant, *p*'s<.01. The partial correlations, when removing the effect of pre-test, are as follows: post-test x negative feedback *r*=.20, post-test x revision *r*=.21, overall poster quality x negative feedback *r*=.22, overall poster quality x revision *r*=.36. The significant pattern of correlations held up for both schools.

Negative feedback and revision were highly correlated. To determine if negative feedback and revision were independent predictors of learning from the game, we entered them simultaneously in two multiple regressions, one using the post-test as the dependent measure and one using overall poster quality. For the post-test, both negative feedback [t(411)=2.7, *p*=.006, β=.15, B=.17, SE=.06] and revision [t(411)=2.8, *p*=.005, β=.16, B=.36, SE=.13] were significant predictors [F(2,411)=16.11, p < .001, R Square=.07, Adjusted R Square=.07]. Similarly, for the overall poster quality score, negative feedback [t(470)=3.2, *p*=.002, β=.15, B=.89, SE=.28] and revision [t(470)=5.4, *p*<.001, β=.26, B=3.23, SE=.59] were significant independent predictors [F(2,470)=36.07, p < .001, R Square=.13, Adjusted R Square=.13].

In sum, two measures of learning — the performance-based poster quality score and the declarative post-test — correlated independently with both the choice to seek more negative feedback and with the choice to revise. This pattern persisted when controlling for prior knowledge. Thus, within the context of the design-oriented *Posterlet*, seeking negative feedback had a positive association with learning, even though both positive and negative feedback were equally informative.

4.2.2   The Amount of Chosen Negative Feedback Positively Correlates with Dwell Time
The standard explanation for the learning benefits of negative feedback is that it provides information for improvement, whereas positive feedback reinforces existing behaviours, not new learning (Hattie & Timperley, 2007). However, a second possibility is that people may also pay more attention to negative

feedback, which could further contribute to learning. For example, Hancock, Stock, and Kulhavy (1992) found that if students aged 15 to 18 were confident of a correct answer on SAT-like questions and they received negative feedback, they spent more time reading that feedback. We assume the students were preparing to take the SAT, so they had high motivation to learn. However, Chase et al. (2009) found that children tended to pay less attention to negative feedback they received in a knowledge construction task, such as building a concept map. *Choosing* to receive negative feedback may mitigate the tendency to under-process (unwanted) negative feedback. To test the relation of negative feedback and processing, we further examined the amount of time the students spent dwelling on the feedback. The expectation was that students would spend more time processing negative feedback than positive feedback.

As a first step, we correlated the feedback-screen dwell time for each poster with the number of negative feedback messages (0–3) the student chose for that poster. The correlations were as follows: $r_{Poster\ 1}=-.02$, $p=.67$; $r_{Poster\ 2}=.20$, $p<.001$; $r_{Poster\ 3}=.21$, $p<.001$. For posters 2 and 3, there was a positive association between the amount of chosen negative feedback and the dwell time, which we interpret as a closer read of the feedback. The lack of correlation for poster 1 may be attributed to the novelty of the environment, so that for poster 1 there was more exploration and more variability: $M_{round1}=19.1$ seconds, SD=17.2; $M_{round2}=14.8$ seconds, SD=8.7, and $M_{round3}=12.9$ seconds, SD=6.7.

The correlation between the dwell time and amount of negative feedback does not stem from differences in message length: there is no statistically significant difference in the length of the negative and positive feedback messages. However, a reasonable explanation is that students who choose more negative feedback were generally more diligent and they read their feedback more carefully and slowly. A different explanation is that the correlation reflects students selectively paying closer attention to their negative feedback than to their positive feedback. We looked at dwell times within subjects to test this hypothesis. We compared the dwell times for each student for rounds 2 and 3. The logic is that, if the correlation is specific to negative feedback, then a student who chose more negative feedback on round 2 than on round 3 should also exhibit longer dwell times for round 2 than for round 3. Similarly, if a student chose more negative feedback on round 3, then dwell times should be longer for round 3. This is exactly what we found. We constructed a variable that computed the amount of negative feedback chosen for round 2 minus the amount of negative feedback chosen for round 3. We also constructed a variable that computed the feedback dwell time for round 2 minus the feedback dwell time for round 3. The correlation of these two variables was significant; $r=.14$, $p=.003$. This indicates that individuals spent more time processing self-selected negative feedback compared to self-selected positive feedback.

The log data permit further refined analyses. For instance, one might examine whether students who received negative feedback about a specific principle (e.g., font too small to read) performed better on that specific principle when they made their next poster. A difficulty with this analysis is that students would often try out completely different designs on their next poster, so that they did not immediately employ what they had learned. More generally, we searched the log files for many of these types of

micro-changes. We were unable to find patterns more telling than the coarse analyses above, which we applied *a priori*.

4.2.3   Seeking Negative Feedback in the Game Correlates with Academic Achievement

Are we only measuring behaviours and correlations that hold in the context of the game? Ideally, we could observe students choosing feedback across many topics and settings and determine whether their choices correlate with learning in those environments. However, prior to incurring the expense of broad generalization research, a prudent first step is to determine whether our assessment shows any relation to other markers of learning. We received standardized English Language Arts and Mathematics achievement tests for a subset of the students. We correlated learning choices collected from the game with these measures. Table 3 shows correlations by school. The association of negative feedback with achievement outcomes is significant, moderate, and similar across the four tests, whereas revision shows more modest and variable correlations. Notably, the correlations of negative feedback are greater here for the achievement outcomes than for the measures of graphic design principles taught in the game (Table 2). This is likely a function of our learning measures, which had not had the benefit of multiple revisions to improve their precision and reliability.

Table 3. Correlations between negative feedback, revision, and outside assessments

Measures	New York City (n=119)		Chicago (n=65)	
	Reading (NYSTP ELA)	Math (NYSTP Math)	Reading (ISAT)	Math (ISAT)
Negative Feedback	.33***	.39***	.41***	.33**
Revision	.08	.28**	.31*	.21

Note: *** $p<.001$, ** $p<.01$, * $p<.05$

The fact that the choice to seek negative feedback holds relatively high correlations across different populations of students and across different standardized achievement measures is a promising indicator that *Posterlet* measures something of importance. This finding is somewhat surprising because *Posterlet* presents a context of design, whereas the achievement tests emphasize demonstrably correct answers. The fact that negative feedback shows stable correlations with achievement measures, whereas revision does not, provides useful divergent validity. It implies that *Posterlet* is measuring two different things, one of which seems more relevant to academic achievement. One value of divergent validity is that it contributes to theory building. For example, if one believes that the correlation of negative feedback, revision, and learning are due to a latent variable, such as students' IQ or growth mindset (Dweck, 2006), then one needs to explain further why negative feedback has a consistent relation to multiple achievement measures, whereas revision does not.

4.2.4   Choices to Seek Negative Feedback Are Contingent on Experience

Assessments of human abilities always raise the question of whether performance differences result from a relatively immutable trait, such as IQ, or whether the abilities are mutable. In the case of *Posterlet*, the question is whether there is a reason to believe that a course of instruction could change

student willingness to seek negative feedback. While we want to believe the answer is yes, it would be reassuring to find evidence to support the idea that feedback seeking is malleable, before engaging in costly instructional research. As a simple exploration, we tested whether the NYC and Chicago schools exhibited different patterns of choices. The NYC students chose negative feedback (M=3.96, SD=2.48) significantly more often than Chicago students (M=3.24, SD=2.14), $t(402)$=–3.2, p<.01, *df* adjusted for unequal variances. There were no appreciable differences in the rates of revision; M_{NYC}=1.1 (SD=1.14), $M_{Chicago}$=1.2 (SD=1.14), $t(442)$=.68, p=.50. Thus, students from the two schools showed strong differences in the willingness to seek negative feedback, but not in the willingness to revise. The schools further differed in post-test performance, with NYC outperforming Chicago on the two checklist questions; $t(391)$=2.4, *SE*=.25, p=.02. We do not know what caused these differences, which could range from differences between the states, the school climates, the curricular implementations, or even the home environments that led parents to send their children to a game-based school. In the meantime, the results suggest that the choices to seek negative and positive feedback are influenced by experience.

5 DISCUSSION, LIMITATIONS, AND CONCLUSIONS

To support research on self-regulated learning, we are developing relatively short, game-based assessments. It is not our initial goal to test broad SRL theories but rather to provide specificity and behavioural evidence to the theorists. Therefore, the assessments probe for specific behaviours in specific contexts, which may be important SRL behaviours. With *Posterlet*, we tested the frequency of students' choices of negative over positive feedback in the context of designing posters. We also tested whether they chose to revise their work given an opportunity.

An important element of our choice-based assessments is that they include new material that students can learn. This way, the assessments can test whether students have developed SRL behaviours that prepare them for future learning. To be clear, *Posterlet* does not teach students to choose negative feedback and revise. Instead, *Posterlet* measures whether students choose negative feedback and revise, and whether they learn graphic design principles as a consequence. This assessment design has the advantage that it both tests for SRL choices and it also determines whether those SRL choices do, in fact, correlate with learning.

The current work presents our initial empirical validation of *Posterlet*. We collected the log files from hundreds of students who completed *Posterlet* as part of a larger evaluation of two game-themed schools. We found that students who chose more negative feedback also learned more graphic design principles as measured by the quality of their posters and their performance on a post-test. Additionally, these correlations were largely unchanged when controlling for the students' incoming knowledge (as revealed by their first poster before feedback). Stated inversely, students who sought positive feedback learned less. We also found that the choice to revise correlated with improved learning according to the same outcome measures. The choices of negative feedback and revision were correlated, so it is difficult to know if the negative feedback drove greater revision. Nevertheless, negative feedback and revision were separable and independent contributors to student learning, so they should not be subsumed

under the same SRL construct.

To our knowledge, this is a first-of-kind demonstration that choosing negative over positive feedback correlates with learning. Moreover, we found that students who chose negative feedback spent more time attending to that feedback. This was not due to a person-level variable, such as diligence or IQ, because the same child would spend more time reading feedback when choosing more negative feedback. This is an additional first-of-kind demonstration. Prior explanations of why negative feedback enhances learning have focused on the error-correction signal. Here, we also found that when people have the opportunity to choose the valence of their feedback, they spend more time processing negative feedback, which, by nearly any cognitive theory, should lead to greater learning (see also Long & Aleven, 2013).

Somewhat surprisingly, choosing negative feedback in the game also exhibited stable correlations with four separate measures of standardized achievement ($.32 < r < .41$). The correlations are of moderate size, but the fact that there are any correlations is somewhat surprising, because *Posterlet* presents a context of design, whereas standardized achievement tests reflect long-term experiences in more academic contexts. It is a matter for future research to determine the bridge of transfer between behaviours in *Posterlet* and academic outcomes. One can imagine several possibilities for how the choice to avoid negative feedback would play out in children's academic achievement. For instance, children may disregard negative feedback to avoid ego threat (Chase et al., 2009) and children may not ask other people to look at their work for fear of negative feedback. Moreover, there may be a broader disposition of which the choice to seek negative feedback is simply one manifestation. There are a variety of possible candidates that range from self-efficacy to growth mindset. However, to apply these theories, it will be important to account for why the choice to revise did not exhibit equally stable correlations across the standardized achievement tests ($.08 < r < .31$). Growth mindset, for instance, should predict both the motivation to gain information for improvement (negative feedback) and the motivation to take action (revise) based on that information. However, only the choice to seek negative feedback exhibited a consistent relation with academic achievement.

The current results reflect initial work and, as such, the findings must be scoped within the typical limitations of early correlational research that has not had a chance to test for replicability, generality, and causality. First, we were not present during the administration of the assessment. The proctors, whom we never met, were blind to our hypotheses and had not seen *Posterlet* previously. Still, it is possible that some unknown events drove the results.

Second, our main findings are that the choice of more negative feedback is associated with greater learning and longer feedback dwell times. It is important to recognize this is a single demonstration with a unique assessment using a single age group. Whether this finding holds in more everyday settings and for other learner profiles has not been demonstrated, and therefore, cannot be claimed (but see Cutumisu & Schwartz, 2014; Cutumisu, Chin, & Schwartz, 2014). The correlations with academic

66

achievement measures suggest there is some relation to other contexts of learning, but the nature of that relation is unknown. This is typical of behavioural research, which depends on replication and testing under a variety of circumstances before claiming generalizability. In the meantime, the finding may serve as a useful data point for SRL theories.

Third, our research was primarily correlational and designed to evaluate our assessment. We cannot make causal claims that choosing more negative feedback is the cause of better learning. As always, there may be latent variables (e.g., persistence) that drove the correlations of negative feedback selection, revision, and learning. Similarly, we cannot make causal claims about how the choice to seek negative feedback in *Posterlet* is related to performance on the standardized achievement measures. However, given the initial demonstration, it is now possible to consider research designs that could pursue questions of causality.

Fourth, we have no way to determine the sources of individual differences in the selection of negative feedback, which could be a function of parental income, school curricula, weather, and many other factors (Aikens & Barbarin, 2008). For example, we found that the NYC students chose more negative feedback than the Chicago students. For all we know, this may be a result of being raised in NYC and may have nothing to do with their respective school experiences. From our perspective, the source of individual differences is of less interest than the question of whether we can instruct students to seek negative feedback, and whether, in turn, this will improve their learning. An experiment on this question is currently underway and we are using *Posterlet* as a post-test to determine whether students transfer their SRL lessons from a classroom setting to the fanciful design world of *Posterlet*.

Thus far, the work on *Posterlet* has been theoretically agnostic. *Posterlet* was not designed to test any broad SRL theory or theory about types of learners (e.g., fixed versus growth mindset). The goal was to create a theory-neutral, process-based behavioural assessment of a potential SRL behaviour and, simultaneously, to determine if this SRL behaviour is useful for learning. Now that this initial work is done, it is possible to begin theoretical work.

One important place to focus that theoretical work is the difference between spontaneously choosing to engage in an SRL behaviour versus being directed to engage in that behaviour. For example, letting patients choose their level of pain medication led to lower doses than when the doses were prescribed by medical staff (Haydon et al., 2011). Similarly, letting students choose negative feedback rather than assigning it to them may lead them to learn more from that feedback. Therefore, we are currently using *Posterlet* to investigate the effect of chosen versus assigned negative feedback. If choosing leads to better learning, it implies that our focus should not simply be on teaching an SRL strategy, but also on convincing students to choose that strategy, so they will reap the most benefit. While this point may seem obvious, one non-intuitive hypothesis is that the benefits of SRL will be more pronounced in transfer settings, where students choose to use those strategies, than in the original instructional context that required students to use those strategies.

In conclusion, students playing *Posterlet* exercise free choice in deciding to seek negative feedback or to revise, these choices being equally acceptable, rather than right or wrong. Assessing student choices provides a new approach for evaluating process skills that have been elusive to more traditional testing, but that are of great interest to many educators (Schwartz & Arena, 2013). Instead of considering choice as self-selection and source of motivation, we view choice as one of the most important outcomes of learning. By capturing students' choices, we approach our goal of measuring students' propensity for SRL. We aim to build a suite of choice-based assessments that can differentiate which educational experiences foster independent learners who can make good choices for learning in the future.

ACKNOWLEDGEMENTS

REFERENCES

Aikens, N. L., & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology, 100*(2), 235–251.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 531–540). Berlin/Heidelberg: Springer.

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287–314.

Baker, R. S., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education*, Vol. 24 (pp. 61–100). Washington, DC: American Educational Research Association (AERA).

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*(3), 245–281.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology*, *18*(4), 334–352.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology*, Vol. 5 (pp. 161–237). Mahwah, NJ: Erlbaum.

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 J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 36–43). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/36_EDM-2014-Full.pdf

Chi, M., VanLehn, K., & Litman, D. (2010). Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 224–234). Berlin/Heidelberg: Springer.

Chin, D. B., Blair, K. P., & Schwartz, D. L. (2015). Got game? A choice-based learning assessment of data visualization skills. Paper presented at *American Educational Research Association Annual Conference*, Chicago, IL, April 16–20.

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

Cutumisu, M., Blair, K. P., Schwartz, D. L., & Chin, D. B. (2015). A game-based assessment of children's choices to seek feedback and to revise. Paper presented at *American Educational Research Association Annual Conference*, Chicago, IL, April 16–20.

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

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. *Proceedings of the London International Conference in Education* (pp. 171–176), London, UK, November 10–12.

Dweck, C. (2006). *Mindset: The new psychology of success*. New York, NY: Random House LLC.

Fancsali, S. E. (2014). Causal discovery with models: Behavior, affect, and learning in cognitive tutor algebra. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 28–35). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/28_EDM-2014-Full.pdf

Finkelstein, S. R., & Fishbach, A. (2012). Tell me what I did wrong: Experts seek and respond to negative feedback. *Journal of Consumer Research*, *39*(1), 22–38.

Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. *Adult Education Quarterly*, *48*(1), 18–33.

Hancock, T. E., Stock, W. A., & Kulhavy, R. W. (1992). Predicting feedback effects from response-certitude estimates. *Bulletin of the Psychonomic Society*, *30*(2), 173–176.

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

Haydon, M. L., Larson, D., Reed, E., Shrivastava, V. K., Preslicka, C. W., & Nageotte, M. P. (2011). Obstetric outcomes and maternal satisfaction in nulliparous women using patient-controlled epidural analgesia. *American Journal of Obstetrics and Gynecology*, *205*(3), 271, e1–6.

Howard-Rose, D., & Winne, P. H. (1993). Measuring component and sets of cognitive processes in self-regulated learning. *Journal of Educational Psychology*, *85*(4), 591–604.

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), 978–992.

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.

Lee, S. J., Liu, Y. E., & Popović, Z. (2014). Learning individual behavior in an educational game: A data-driven approach. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 114–121). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/114_EDM-2014-Full.pdf

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 75–105). Hillsdale, NJ: Erlbaum.

Long, Y., & Aleven, V. (2013). Skill diaries: Improve student learning in an intelligent tutoring system with periodic self-assessment. *Artificial Intelligence in Education: Lecture Notes in Computer Science, 7926*, 249-258. Berlin/Heidelberg: Springer.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology*, 2 (pp. 745–783). New York, NY: Taylor & Francis.

Okita, S. Y., & Schwartz, D. L. (2013). Learning by teaching human pupils and teachable agents: The importance of recursive feedback. *Journal of the Learning Sciences*, *22*(3), 375–412.

Peacock, S. B., Smith, C. P., Martin, T., Aghababyan, A., Popović, Z., Andersen, E., & Liu, Y. E. (2013). Learning fractions through splitting in an online game. *Proceedings of the American Educational Research Association Annual Conference (AERA)*, San Francisco, CA, April 29. Retrieved from http://www.aera.net/Publications/OnlinePaperRepository/AERAOnlinePaperRepository/tabid/12720/Owner/604309/Default.aspx

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40.

Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review, 85*, 417–435.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011a). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011b). Metacognitive practice makes perfect: Improving students' self-assessment skills with an intelligent tutoring system. In *Artificial Intelligence in Education: Lecture Notes in Computer Science, 6738,* 288–295. Berlin/Heidelberg: Springer.

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 Scienc*es, 23, 537–560. doi: 10.1080/10508406.2014.883977

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

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 352–359.

Schooler, L. J., & Anderson, J. R. (1990). The disruptive potential of immediate feedback. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 702–708), Cambridge, MA.

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

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for learning: The hidden efficiency of original student production in statistics instruction. *Cognition and Instruction, 22*, 129–184.

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 U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Abingdon, UK: Routledge.

Snow, E. L., Allen, L. K., Russell, D. G., & McNamara, D. S. (2014). Who's in control? Categorizing nuanced patterns of behaviors within a game-based intelligent tutoring system. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 185–191). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/185_EDM-2014-Full.pdf

Stevens, R. H., & Thadani, V. (2007). Quantifying student's scientific problem solving efficiency and effectiveness. *Technology, Instruction, Cognition and Learning*, *5*(4), 325–337.

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

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572.

Zeidner, M., Boekaerts, M., & Pintrich, P. R. (2000). Self-regulation: Directions and challenges for future research. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 749–768*)*. San Diego, CA: Academic Press.

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