British Journal of Educational Technology *Vol 50 No 6 2019 2977-3003*
doi:10.1111/bjet.12796

A Digital Game-based Assessment of Middle-School and College Students' Choices to Seek Critical Feedback and to Revise

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

cutumisu@ualberta.ca, dbchin@stanford.edu, danls@stanford.edu

[1]Department of Educational Psychology, University of Alberta, Edmonton, Canada

[2]Stanford Graduate School of Education, Stanford, US

Corresponding author:

Dr. Maria Cutumisu, 6-102 Education North Centre, Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, T6G 2G5, Canada, Telephone: 780 492 5211, Fax: 780 492 1318, Email: cutumisu@ualberta.ca.

Maria Cutumisu is an assistant professor in the Department of Educational Psychology at the University of Alberta. She received M.Sc. and Ph.D. degrees in Computing Science from the University of Alberta and she trained as a postdoctoral scholar in Learning Sciences at the Stanford Graduate School of Education. Her research interests include game-based assessments, artificial intelligence in games, feedback processing, data science, and eye tracking.

Doris B. Chin is a senior research scholar with the Graduate School of Education at Stanford University. As a former science museum educator, she is passionate about exploring innovative and engaging ways to help people learn, in both formal and informal learning environments. Her research has a special focus on STEM concepts, strategies, and practices.

Daniel L. Schwartz is the Dean of the Stanford University Graduate School of Education and holds the Nomellini-Olivier Chair in Educational Technology. He is a learning scientist who also spent eight years teaching secondary school in Los Angeles and Kaltag, Alaska. His special niche is the ability to produce novel and effective learning activities that also test basic hypotheses about how people learn.

Citation:

Cutumisu, M., Chin, D. B., & Schwartz, D. L. (2019). A digital game-based assessment of middle-school and college students' choices to seek critical feedback and to revise. *British Journal of Educational Technology, 50*(6), 2977-3003. https://doi.org/10.1111/bjet.12796

Abstract

A major goal of contemporary education is to teach students how to learn on their own. Assessments have largely lagged behind this goal, because they measure what students have learned and not necessarily their learning processes. This research presents Posterlet, an assessment that collects evidence regarding the strategies that students choose while learning on their own. Posterlet is an educational game-based assessment that measures two design-thinking choices: students' choices to seek critical (i.e., negative) feedback and to revise their work while they learn graphic design principles through creating posters. This research also presents an examination of students' choices to seek feedback and to revise, as well as of students' learning outcomes based on these choices. This game-based assessment approach is empirically validated with three research studies sampling nearly 300 middle-school and college students who played Posterlet and completed a post-test. Results show that the game helps students learn, as students who play the game before completing the post-test learn more graphic design principles than students who only complete the post-test. Moreover, the choices to seek critical feedback and to revise can predict learning and can be used as valid outcome measures for learning. Findings can be used in developing and evaluating models of instruction and assessment that may help students make informed learning choices. A discussion of present and future trends in theory regarding digital feedback environments is also included.

Keywords

feedback seeking; learning; performance; revision; game-based assessment; artificial intelligence

Practitioner Notes

What is already known about this topic:

- The link between critical feedback and learning outcomes is still a matter of debate.
- Most research focuses on feedback that is assigned to the students rather than chosen.
- Research and theory on game-based learning are in their early stages.

What this paper adds:

- Posterlet is an assessment game that helps students learn graphic design principles.
- Students' choices to seek critical feedback and to revise can predict learning.
- Learning choices can be used as valid outcome measures for learning.
- Students' choices within the game are positively associated with students' learning experiences outside of our assessment environment (e.g., in school).

Implications for practice and/or policy:

- Findings can be used in developing and evaluating models of instruction and assessment that may help students make informed learning choices.
- Short assessments like Posterlet can measure student behaviors that are more predictive of a range of learning outcomes and future performance than testing a narrow knowledge set that may soon become obsolete.
- Such assessments could also provide formative feedback to educators to help them improve their programs of instruction.

Introduction

One of the most important roles of education is to facilitate autonomous learners who are able to innovate in the ever-changing environment of the 21st century (Piaget, 1964; Yackel & Cobb, 1996). One way to know if we succeeded in our endeavor is to use the right assessment to measure students' potential to innovate.

However, most traditional assessments do not offer individuals opportunities to learn during the test. In contrast, dynamic assessments are interactive assessments used in education, stemming from Lev Vygotsky's social development theory of learning and intended to measure the *zone of* *proximal development* (Vygotsky, 1997). They are alternative assessments that aim to measure students' learning processes and to identify not only the current level of understanding of individuals but also their potential to learn in the future. For instance, such assessments can help determine whether an individual has the potential to learn a new skill, such as a new language. The difference between a traditional (or static) assessment and a dynamic assessment is that the former evaluates the performance on the given test materials while the latter evaluates the performance response to an intervention in which the examiner first instructs the examinee on how to perform better (Sternberg & Grigorenko, 2002).

Digital Feedback Environments: Present and Future Trends in Theory

Artificial intelligence and machine learning techniques have been increasingly used to create adaptive learning environments (e.g., intelligent tutoring systems) that help students learn and provide customized feedback using data mining and learning analytics (Aleven & Koedinger, 2000; Roll, Aleven, McLaren, & Koedinger, 2011; Snow, Varner, Russell, & McNamara, 2014). It is largely thought that, even in these environments, learning during assessment is rarely possible. Indeed, research and theory on game-based learning are in their early stages and the theoretical mechanisms underlying game-based learning are still not understood (Mayer, 2018). A recent review of scientific research on computer games for learning evaluated the research effectiveness and progress of game research. Although it found cognitive gains when training with computer games (e.g., using first-person shooter games to train perceptual attention skills and using spatial puzzle games to train two-dimensional mental rotation skills) as compared to conventional approaches, it also highlighted mixed results regarding learning gains in these environments, recommending further exploration of the cognitive, motivational, affective, and social processes that underlie learning with educational computer games (Mayer, 2018). A recent meta-analysis confirmed that more research is needed to clarify which instructional techniques are effective for learning in serious games (i.e., computer games used for learning, training, and instruction; Wouters & van Oostendorp, 2017). Despite this, recent research findings suggest that even short games may be effective for learning. Most notably, researchers developed a short 30-minute "fake news game" in which n = 95 high-school students deliberately created a fake news article (Roozenbeek & van der Linden, 2018). The randomized study conducted in a public-school setting showed that playing the fake news game "inoculated" players against fake news, reducing the perceived reliability and persuasiveness of fake news articles.

In the context of the advent of technology and psychophysiological methods, learning theorists have posited that psychomotor activities (e.g., motion or body movements) are integral parts of cognition and learning. Specifically, they highlight current trends in learning sciences that emphasize the role of advanced motion capture technologies in providing advanced feedback to support learning through a symbiotic relation between body and sensorimotor engagement (Starčič, Lipsmeyer, & Lin, 2019). The mechanism through which motion capture technologies support learning include fostering a seamless human-computer interaction and a more direct link between physical motion and psychological reactions, and bridging enactive learning with model-based learning. This trend will complement computer-based learning, which mainly supports visual and symbolic representations of information or knowledge, by extending it to enactive representations and by unifying the different types of representations to improve learning through the provision of adaptive feedback. It will also enhance existing efforts to collect psychophysiological data, such as electrodermal activity, eye movement, and brain wave patterns.

Even in such success cases, as mentioned before, traditional dynamic assessments embed some type of an instructional intervention within the assessment procedure. They usually involve direct instruction as the way students learned during the assessment. This approach is effective, as a recent meta-analysis revealed that serious games with instructional techniques improve learning more than serious games without instructional techniques, with the weighted mean effect size of learning for instructional techniques being $d = .41$ ($z = 7.29$, $p < .001$; Wouters & van Oostendorp, 2017). Some instructional techniques (e.g., modeling the essential information for solving a problem and how to solve a problem: $d = .55$, $p < .001$, feedback: $d = .63$, $p < .001$) improve learning significantly relative to serious games without these techniques. Thus, in most dynamic game assessments, players have no or limited control of the game from an instructional perspective (Graesser, 2017).

To address these issues related to the measurement of students' potential to learn and to innovate, we have been building a suite of online games, called choicelets, designed as interactive assessments of 21st-century learning skills, including choosing to seek feedback and to revise (Cutumisu, 2018; Cutumisu, Blair, Chin, & Schwartz, 2015, 2017; Cutumisu, Chin, & Schwartz, 2014; Cutumisu & Schwartz, 2018), and design-thinking skills (Chin, Blair, Wolf, Conlin, Cutumisu, Pfaffman, & Schwartz, 2019; Conlin, Chin, Blair, Cutumisu, & Schwartz, 2015). Recently, these games were also used to assess the role of fixed and growth mindset onto university students' willingness to seek critical feedback and to revise (Cutumisu, 2019). These games log students' actions with the aim of gaining an insight into students' learning processes, specifically their choices, as they attempt to solve challenges. Such games could also help further identify behaviors that are critical for learning. They are short (e.g., 15-30 minutes of gameplay), stand-alone assessments that are not designed to target a specific curriculum, but rather to evaluate a range of learning experiences. As they explicitly use information from the game environment to infer players' capabilities, these games are situated within the umbrella of game-based assessments (Mislevy et al., 2014). Some of the use cases for game-based assessment include collecting information about players' actions in the game to adapt to the players or to provide feedback to aid learning, embedding specific learning outcomes, providing formative assessment opportunities for students or their teachers, while also collecting gameplay information that can help designers improve their games.

This research hypothesizes that the choices students make when presented with a challenge constitute important behaviors for learning. Also, it describes *choice-based assessments* that measure not only students' knowledge but, even more importantly, their choices about what, when, and how to learn. Measuring learners' behaviours through learning analytics as they engage with feedback has focused on self-reported data collected through surveys and questionnaires (Mahfoodh, 2017; Mulliner & Tucker, 2017; Narciss, 2013; Sargeant et al., 2011). In contrast to observational or survey methods, assessment games that collect students' choices can provide information about effects beyond the observed learning experience. Finally, this research empirically validates the Posterlet game-based assessment environment, an instance of choice-based assessments, with three research studies showing that choices can both predict learning and can also be used as valid outcome measures for learning.

Thus, in contrast to traditional assessments in which there are no opportunities to learn during the assessment, the Posterlet assessment is designed to offer players several learning opportunities. Our first research question asks whether students learn during the assessment in Posterlet. In Posterlet, we are also studying an interesting context of learning that differs from traditional dynamic assessments that usually involve direct instruction as the mechanism through which students learned during the assessment. Instead, we are examining the choices people make during learning as the mechanism through which students learn during the assessment. Specifically, instead of direct instruction, Posterlet offers players a choice of the type of feedback they receive, which enhances their sense of autonomy. Players choose between critical (negative) and confirmatory (positive) feedback from game characters, and they also choose whether to revise their digital posters. This leads to our second main research question: which choices seem to drive learning? In other words, do students indirectly learn from the consequences of their choices (e.g., the feedback these characters provide)? Finally, our third research question is whether learning in the game provides any evidence regarding learning in the "real world" and whether the answers to the previous two research questions can generalize to other populations. Thus, the question examines external validity of learning choices.

Theoretical Framework

This research is grounded in the theoretical framework of *constructivist assessments* (Schwartz, Lindgren, & Lewis, 2009) and in *choice-based assessments* (Schwartz & Arena, 2009, 2013). The former constitutes a class of assessments embedding opportunities for students to learn during the assessment (e.g., students learn graphic design principles from feedback while designing posters and being assessed on these principles). The latter represents a type of assessment that measures the choices students make while they solve a challenge (e.g., the strategies students employ to design better posters, such as choosing critical feedback and revising their posters). Both these assessments measure not only learners' knowledge outcomes but also, more importantly, their learning processes (e.g., choices about what, when, and how to learn). Choices constitute the core of the current assessment approach, being considered to be first-order learning outcomes and measurement constructs. Assessment environments that enable the tracking of students' choices and learning outcomes are needed, so that researchers can examine the impact of choices on learning trajectories and measure students' potential to learn independently. Ideal assessment environments that have the potential to measure students' preparedness to be independent learners must satisfy the following three principles: *typical performance*, *preparation for future learning*, and *choice*.

Principles of Choice-Based Assessment Environments

Typical Performance

Assessments need to capture every-day learning behaviors. In traditional assessments, students are expected to display "maximal performance" or test behavior during the assessment (Klehe & Anderson, 2007; Sackett, Zedeck, & Fogli, 1988), which may not be compatible with learners' outside-school behaviors and performance. Instead, assessment environments need to create an atmosphere in which students feel comfortable experimenting with different strategies and displaying their typical problem-solving behaviors. Such settings may enable the evaluation of students' true trajectories for lifelong learning and may predict students' behaviors outside school, where they need to make good learning choices on their own. Posterlet is designed to facilitate students' unconstrained typical behaviors (e.g., their attitudes and approaches towards choosing critical feedback and revising), while they learn about graphic design principles (e.g., acquired proficiency and knowledge about poster design). Assessments such as Posterlet have the potential to pave the way towards a comprehensive competency model of an individual, by inspecting not only individuals' knowledge performance, or hard skills, but also their more elusive characteristics, or soft skills, that are predictive of students' learning and performance in new contexts. Together, these models could prepare individuals to become independent and adaptable learners to a wide variety of veridical learning situations.

Preparation for Future Learning (PFL)

Assessments need to offer learning opportunities. Most assessments are retrospective, measuring students' knowledge at the end of instruction. Thus, they offer a snapshot of a current state of students' accumulated knowledge, but provide little indication regarding students' future learning trajectories. According to Vygotsky (1997), measuring student knowledge at the end of instruction, instead of measuring learning processes progressively, does not yield an accurate, comprehensive view of students' learning growth and development. Thus, if the goal of assessments is to measure behaviors that are conducive to learning, then assessment environments need to provide learning opportunities. Researchers proposed *preparation for future learning* (PFL) assessments that afford such opportunities for students to learn during their evaluation (Bransford & Schwartz, 1999). They are prospective assessments, as they aim to measure how well students are prepared to learn in the future, rather than just how well prepared they are at the time of the assessment. Posterlet is an example of a PFL assessment in which students have the opportunity to learn a subset of the 21 graphic design principles through seeking confirmatory (i.e., positive) or critical (i.e., negative) feedback in the game. It is important to note that critical feedback here means constructive feedback and not punishment, and that both feedback valences (i.e., confirmatory and critical) provide equivalent informational value and equal learning opportunities. It is also important to note that Posterlet was intended to be an assessment of students' learning choices rather than an instructional game that teaches students to design better posters. We employed a relatively unknown learning task, poster design, to provide learning opportunities in the assessment environment. This way, the effect of students' choices on their learning outcomes can be measured.

Choice

Assessments need to gather information about unforced student choices. Most large-scale assessments consist of multiple-choice test items that provide choice over the answer items testing domain knowledge, rather than over the applicable learning strategies that could help solve many different problems. However, we need to emulate the choice-rich environment in which students will learn after they leave school, because lifelong learning is based on free choices. Some assessment environments blend learning and assessment, such as intelligent tutoring systems (Koedinger, Anderson, Hadley, & Mark, 1997). In contrast to our approach that emphasizes the value of students' choices, these environments assume a certain sequence of steps that students take while learning, giving little actual choice to the students. An important feature of behavioral assessments is that they cannot direct or influence students' choices. For instance, in a game-based assessment, students should be able to level up, no matter what learning choices they make in the game. In Posterlet, students have the same opportunities to learn and to advance through the game rounds, regardless of their choices between confirmatory or critical feedback, and of their choices to revise their posters. Each student has equal opportunities to choose either confirmatory or critical feedback from each of the three virtual animal characters on each game round. Similarly, students have equal opportunities to revise, or not, their poster on each game round. In summary, students can complete each digital poster regardless of the choices exercised throughout the game.

The Measurement Constructs

The Posterlet online game is an instance of a choice-based PFL assessment designed to collect and measure two behaviors important for learning, namely students' choice to seek critical feedback from interactive game characters and students' choice to revise posters. This assessment game is typical of a graphic design environment, which involves an artificial learning task of designing a digital poster. This creative, open-ended task enables researchers to control variables and to measure students' feedback-seeking choices, thus, their preparedness to learn on their own. It is more representative of most tasks and assessment that students encounter in real learning environments than in school, and it presents little variation in learners' prior experience.

Feedback is an important ingredient in learning a new task, because it constitutes information about a person's performance or understanding (Black, 2015; Hattie & Timperley, 2007). The feedback literature highlights the importance of feedback for learning and motivation (Ammons, 1956; Ilgen, Fisher, & Taylor, 1979; Mory, 2004). For instance, a synthesis of over 800 meta-analyses related to student achievement revealed a 0.75 overall effect size for feedback (Hattie, 2009), while a more recent review of 12 meta-analyses of feedback (Hattie, 2013) revealed an average effect size of 0.79. These large effect sizes quantify the magnitude of the difference in student achievement between a control group and a feedback intervention experimental group, showing the effectiveness of the feedback intervention in a wide range of contexts. For example, an effect size of 0.79 indicates that the score of the average student in the feedback intervention group is 0.79 standard deviations above the score of the average student in the control group. Informative feedback is crucial in developing mastery in many domains, especially when it is accompanied by revision (Ericsson, Krampe, & Tesch-Römer, 1993). Despite this, the specific mechanisms relating feedback to learning are still not well understood (Attali, 2015). The feedback literature also warns about its deleterious effects on performance in some situations (Kulik & Kulik, 1988). In a meta-analysis of feedback, Kluger and DeNisi (1998) found that feedback improved performance by an average of 0.4 standard deviations compared to no feedback controls. However, in a third of the studies, feedback hindered performance (i.e., feedback was worse than no feedback at all).

Generally, in most research studies, feedback is assigned to the student by an instructor or by a computer. In contrast, this research examines the situation in which students had a choice regarding the valence of their feedback. It focuses on the choice to seek critical feedback, because critical (i.e., negative) feedback tends to be more effective for continued performance than confirmatory (i.e., positive) feedback (Kluger & DeNisi, 1998). At the same time, critical feedback runs the risk of triggering an ego threat that leads people to disregard critical feedback rather than engage with it (Hattie & Timperley, 2007). While attitudes towards feedback are important for learning, there is no evidence supporting that the choice to seek feedback is important. Revising may also be an important learning behavior, as choosing critical feedback would be of little use if students did not act on it. Although revising seems like an important behavior for learning, there is no evidence in prior research whether the choice to revise is important as well. Moreover, research shows that feedback information is rarely used in revision of work (Carless, 2006). Thus, the impact of the *choices to seek critical feedback* and *to revise* on learning is investigated.

The Assessment Instrument: Posterlet

Players design posters for different booths at a fun fair in Posterlet. In the short introduction provided by the game, the players learn that they were selected to join the school's committee in charge of designing posters for three different booths at the school's Fall Fun Fair (Step 1). Then, in Step 2 (*Design Poster*), students design posters by selecting text and images from a library, placing them on the poster canvas, and adjusting their appearance. In Step 3 (*Choose Focus Group*), on each game round, after they complete their initial poster design, players can ask a focus group of virtual animal characters their opinion about their poster. Thus, players choose three virtual animal characters from the game's focus group of interactive characters. These characters provide more interactivity to the game and remove the stigma of a real social interaction associated with seeking critical feedback from individuals. In Step 4 (*Choose Negative or Positive Feedback*), from each character, players can choose either confirmatory feedback (e.g., "Your poster has big letters. Really easy to read.") or critical feedback (e.g., "People need to be able to read it. Some of your words are too small."), which carry equivalent information, as shown in Figure 1. Posterlet tracks the amount of critical feedback sought by each student on each poster. Then, after Step 5 (*Read Feedback*) in which they read the feedback, players choose whether to revise (Step 6, *Redesign Poster*) or submit their poster (Step 7, *Post Poster and See Ticket Sales*). Posterlet also tracks the number of posters each student revised. There is one round of feedback and revision for each poster. The game has three rounds. Thus, players have nine feedback choices (3 choices to seek feedback per poster x 3 posters) and three revision choices (1 choice to revise per poster x 3 posters). Finally, the game displays students' poster score as the number of tickets sold at each booth. The poster booth with the highest number of tickets sold will bring the most revenue to the school at the Fall Fun Fair.

Figure 1. *After designing a poster, a player has just chosen confirmatory feedback from the elephant, critical feedback from the lion, and ponders whether to choose confirmatory or critical feedback from the panda. Reprinted from Cutumisu, Blair, Chin, and Schwartz (2015).*

The procedure for computing the number of tickets sold is based on the number of graphic design rules used correctly and incorrectly, respectively, on each poster and not on the students' artistic flair (Cutumisu, Blair, Chin, & Schwartz, 2015, 2017). The game's graphic analysis system tracks the posters automatically for the use of 21 design principles commonly used in graphic design. These principles were organized into three broad categories: readability (e.g., one principle posits that there needs to be a high contrast between the color of the text and the colour of the poster canvas so that the text is legible, another that the text must be large enough to be read from a distance), crucial information (e.g., the audience must be provided with crucial information regarding the event, such as the day or the location of the fair), and spacing (e.g., the text or graphics should be spaced out and not touch the edge of the poster). Posterlet uses these principles to assess the quality of an open-ended poster design task. It also includes an intelligent feedback system that provides adaptive feedback based on a priority scheme that ensures a balanced coverage of all 21 rules. Thus, every feedback message is selected successively from one of the three categories (readability, crucial information, and spacing) based on a rule-based artificial intelligent system described in detail in prior research (Cutumisu, Blair, Chin, & Schwartz, 2017). This system generates poster-specific feedback delivered via the virtual animal game characters of the game's focus group. The principles governing choice-based assessments (typical performance, PFL, and choice) are naturally built into Posterlet's game mechanics, providing an environment that facilitates students' *typical behaviors*, offering them opportunities to *learn graphic design principles* while exercising their *feedback and revision choices*. Concomitantly, Posterlet provides researchers with a way to track players' behaviors and learning outcomes with the aim to infer how prepared players are to learn on their own in new learning situations.

Experimental Overview and Study Design

Three studies employing Posterlet as a behavioral assessment instrument were designed to gather evidence on the validity of student choice as a learning construct. Study 1 and Study 2 probe the internal reliability of the Posterlet assessment instrument. Study 3 investigates the external validity of the assessment.

Study 1 was designed to be both correlational and experimental to investigate whether college students learn from playing the Posterlet assessment game, thus testing the hypothesis that Posterlet is a dynamic assessment. In contrast to traditional assessments in which there are no opportunities to learn during the assessment, this study sets out to establish whether there is viable learning in the assessment environment by testing a direct causal claim that playing the Posterlet assessment game improves learning. Specifically, this study compared college students who did not play Posterlet (Non-players or Control condition) to college students who played Posterlet (Players or Treatment condition) to gauge whether playing the game helped students learn design principles. This study ultimately probes the second principle of choice-based assessments, *Preparation for Future Learning* (i.e., learning during the assessment). Without evidence of learning in Posterlet, we cannot relate students' choices to their learning, thus, we cannot know whether some choices students make in Posterlet are beneficial for their learning within this assessment environment.

Study 2 was correlational and it investigated whether the choices to seek critical feedback and to revise correlated with internal learning outcomes for college students. Hence, this study focused on learning choice and performance measurements obtained from the Posterlet assessment environment and learning outcomes obtained from an independent follow-up assessment of students' learning of the graphic design principles encountered in Posterlet. Additionally, Study 2 investigated whether the findings of Study 1 could be replicated even when the learning measures have changed. Thus, this study investigated the internal consistency/reliability of the Posterlet instrument, specifically, whether the new learning measure probed the same construct (i.e., graphic design principles). This study also scrutinized the correlations between choices and learning outcomes and compared the patterns of results to those of Study 1.

Study 3 was correlational and it was designed to: 1) investigate whether the findings of Study 1 and Study 2 can generalize to other ages by sampling a different population, middle-school students; and 2) examine external validity through correlations between choices and external learning outcomes (i.e., in-school standardized achievement scores not related to the Posterlet assessment environment). In other words, our third research question is whether learning in the game provides any evidence regarding learning in the "real world".

Construct Validation

The Posterlet assessment environment is designed to track and measure the free learning choices that students make while they learn how to design digital posters. To validate choices as outcomes of learning, one needs to establish whether some choices (e.g., seeking critical feedback or revising) are better for learning than others. However, in such novel, choice-based assessments, it is difficult to decide whether choices are equivalent or whether one choice is better than another. Thus, to compare the effectiveness of two different learning choices, choice-based PFL assessments face a challenge that knowledge-based assessments do not encounter. Specifically, in assessments of knowledge, the correct answer is known (e.g., 2 + 2 equals 4 and not 5). As there is no literature available on this topic, the major goal of the current research is to show that some choices are better than others for learning. This research employs different forms of evidence to examine two aspects of validity: internal validity (i.e., in the Posterlet assessment environment) and external validity (i.e., outside of the Posterlet assessment environment).

Internal Validity

The Posterlet game assessment was used to gather evidence on the internal validity of choice as a learning construct. Posterlet collects a wealth of detailed log data. However, this research only focuses on two specific types of data: internal learning choices and internal learning outcomes. The major question asked to probe Posterlet's internal validity was the following: "Do choices to seek critical feedback and to revise correlate with *in-game* learning outcomes?" Several subsets of learning measures were employed across three different studies to answer this research question.

Internal Learning Choices. Critical Feedback measures the number of times that a student chose critical feedback (i.e., selected the "I don't like" box) throughout the game out of the total feedback opportunities (three feedback choices for each of the posters). In some samples of the population included in these studies, students designed two posters and in others they designed three posters. As such, the range of this measure varied from zero to six in the former case (i.e., two posters) and from zero to nine in the latter case (i.e., three posters). The amount of confirmatory feedback chosen by each student across posters is complementary to the amount of critical feedback chosen. Data regarding students' choices to revise were also collected via Posterlet. As well, Revision measures the number of times that a student chose to revise throughout the game out of the total revising opportunities. Students have one opportunity to revise their work per poster. The range of this measure varied from zero to two, when the task required students to design two posters, and from zero to three, when the task required students to design three posters.

Internal Learning Outcomes. This research distinguishes two classes of internal learning outcome measures in the current overarching assessment environment: in-game measures collected via the Posterlet game to quantify students' poster performance and out-of-game measures collected via the post-test that varies with each of the three studies. The game's graphic analysis system generated a measure representing the quality of the posters created by the students in Posterlet. It was displayed for each poster as the number of tickets sold by their booth. A post-test was included in each study as an independent measure of how many of the graphic design principles each student learned. Data were collected for several learning measures stemming from the post-test, depending on the particular study.

Performance. *Performance* represents an in-game category of measures that evaluates students' poster design skills. *Poster Quality* was employed to operationalize students' overall poster performance across all posters. The Posterlet game's graphic analysis system evaluates each poster by first scoring each of the 21 design principles with 1 if used correctly, 0 if not applicable, and -1 if used incorrectly on that poster, producing an individual poster score. *Poster Quality* sums the individual poster scores of the last poster version on each game round (i.e., either the revised poster or the initial design, if the poster was not revised). Thus, *Poster Quality* ranges from -63 to 63 for three posters and from -42 to 42 for two posters designed, respectively.

Critique. *Critique* represents a post-test category of measures for students' ability to judge posters. A post-test included two open-text questions: the *Common Mistakes* question asked students to provide some common mistakes a poster design novice might make, while the *Written Feedback* question asked students to provide written feedback on a sample poster shown in Figure 2. Each answer was scored by counting the number of graphic design principles that the student had specified in the open-text answer. Thus, each answer score ranges from 0 to 21.

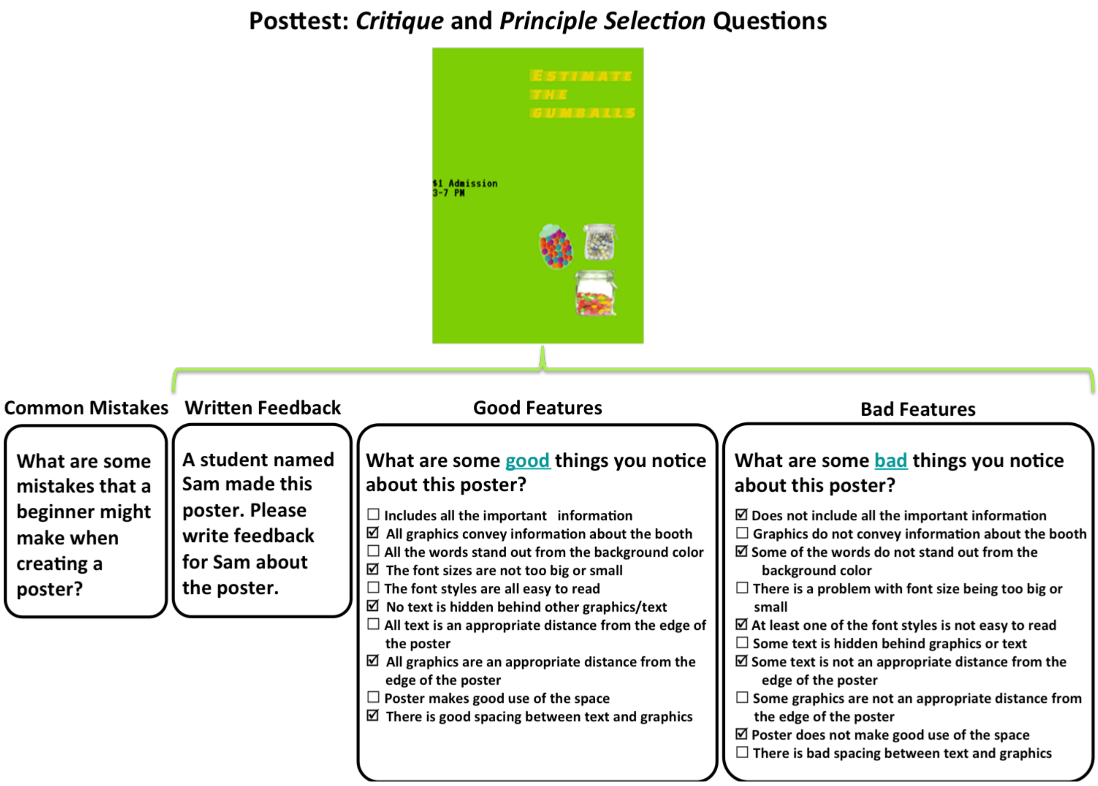

Figure 2. The Critique post-test comprises the Common Mistakes and the Written Feedback questions. Principle Selection comprises the Good Features and the Bad Feature questions that are complementary. The items checked represent the correct answers for the Good Features and Bad Features questions, respectively.

Principle Selection represents a post-test category of measures for a student's understanding of graphic design principles and it is shown in Figure 2. It includes two items, *Good Features* and *Bad Features*, which asked the student to choose, from a checklist of graphic design principles, the aspects that the student deemed to be good and bad, respectively, about a sample poster. Answers were scored by assigning 1 point for each correctly checked answer and by subtracting 1 point for each incorrectly checked answer. Thus, each answer score ranges from -5 to 5, because each question has five correct and five incorrect answers. *Recognition* represents a post-test category of measures quantifying a student's ability to recognize misused graphic design principles at a glance. The post-test included four sets of images and related questions. For each set, students were shown successively three images: a poster, then a distractor Moiré pattern image, and then a second poster, each image being displayed for 5 seconds, as illustrated in Figure 3.

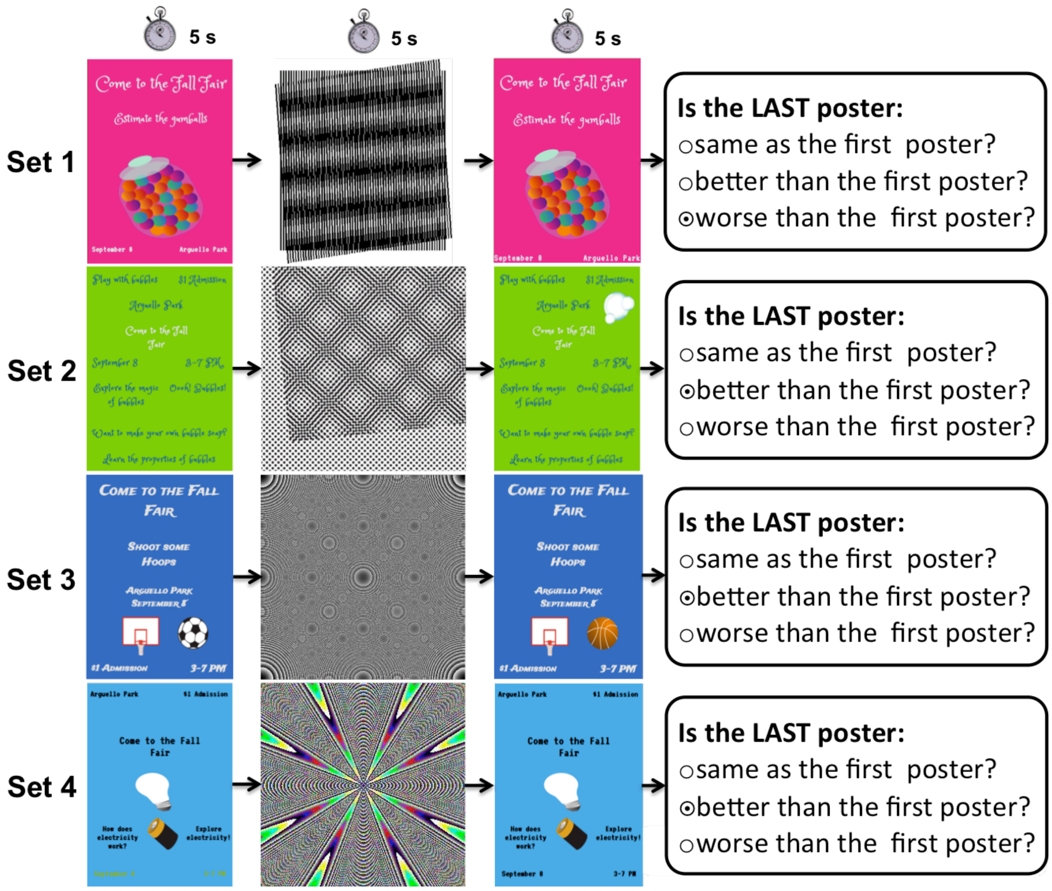

Figure 3. Recognition post-test: students compared the first and the last poster in each of the four sets. Each set targeted a specific graphic design rule: Text not on Edge, Image Present, Graphics Relevant, and Contrast High, respectively.

Students' task was to judge whether the quality of the second poster was the same, better, or worse compared to that of the first poster. Then, students rated their confidence in their answer on a scale of 1 (low confidence) to 5 (high confidence). Finally, they were required to provide a brief written explanation for their decision. A distractor image was inserted between the two posters to ensure that students do not memorize the first poster. This step was taken because previous research demonstrated that humans display an exceptional memory for large numbers of images (Standing, 1973). The current post-test was designed to target the following four out of the 21 design rules: the text should not be situated on the poster's edge, the poster should include at least one image, the images included in a poster should be relevant to the poster's theme, and the text background color contrast should be high (i.e., the text should be easy to read against the background). The post-test evaluates students on their ability to judge posters (i.e., to decide whether the second poster was the same, better, or worse compared to the first poster). The *Poster Ranking* measure was employed to operationalize students' ability to judge posters or to recognize graphic design principles used correctly or incorrectly on a poster. The answer that compared the first and the second poster ("same, better, or worse") in each of the four sets of questions was scored by assigning 1 for a correct and -1 for an incorrect answer (measured by *Text not on Edge, Image Present, Graphics Relevant,* and *Contrast High,* respectively). Thus, *Poster Ranking* represents the number of correct answers for each of the four poster comparisons. It ranges from zero (i.e., students did not judge any of the four poster comparisons correctly) to four (i.e., students judged all four poster comparisons correctly).

External Validity

A post-test that differed from Posterlet immediately followed the game in an attempt to measure the learners' domain knowledge accumulated as a result of playing the Posterlet game. The following question was posed: "Do students who seek critical feedback and revise exhibit better *in-school* learning?" Academic achievement scores were employed in Study 3 to answer this research question.

External Learning Outcomes. Academic Achievement represents an in-school category of measures for middle-school students' English Language Arts (ELA), Mathematics (Math), and Science (Science) standardized test achievement. These tests represent the Standardized Testing and Reporting (STAR) California Standards Tests (CSTs). They were administered for the last time the year before the current study.

Data Analytic Plan

Statistical techniques employed to analyze the data collected from these three studies include tests of association such as correlations and standard linear regressions to determine whether learning choices predict learning and performance outcomes and to explore the associations between the predictors. First, correlation analyses were conducted between the choice and learning measures to ascertain the strength of the relations among the main measures considered in these studies. Then, stepwise linear regressions were conducted to assess whether students' learning choices predicted their learning outcomes together and individually. Lastly, partial correlations were computed between choices and learning measures to gauge the individual contribution of each choice to the relation between the other choice and each learning outcome. Analyses also included tests to compare outcome differences within students (repeated measures analyses of variance) from one poster to another and between groups of students (*t*-tests), such as between Posterlet players (experimental condition) and Posterlet non-players (control condition).

Study 1

Participants, Procedure, and Design

Participants were n = 93 community-college students (70 females, 23 males), aged 18-52, M_{age} = 23.25 years (SD = 6.40). Students were recruited through a student research subject pool and participated in exchange for community college course credit. They provided online consent to participate in the study. As mentioned previously, in addition to being correlational, this study was also experimental. It investigated whether students who played Posterlet before taking the post-test learned more than students who only took the post-test (i.e., whether the Posterlet game helped students learn the graphical design principles). As such, participants were randomly assigned to one of two conditions, control (Posterlet non-players) and treatment (Posterlet players). In the control condition, n = 41 students (35 females, 6 males), aged 18-46, M_{age} = 24.10 years (SD = 7.02), took an online individual post-test (M = 9.08 minutes). In the treatment condition, n = 52 students (35 females, 17 males), aged 18-52, M_{age} = 22.58 years (SD = 5.85), played the Posterlet game individually (M = 8.7 minutes).

Instruments and Measures: The Posterlet Game and the Post-test

The Posterlet game version used in this study required students to design only two posters and it also limited students' time to five minutes per poster. Any potential poster revision was also limited to five minutes. Then, students completed the same individual online post-test (M = 6.63 minutes) as their peers in the control condition. Students were explained the post-test task in the introductory paragraph of the post-test. Then, they were presented with a tutorial image set (illustrated in Figure 4) and a trial image set (illustrated in Figure 5), followed by the main post-test question sets (illustrated in Figure 3).

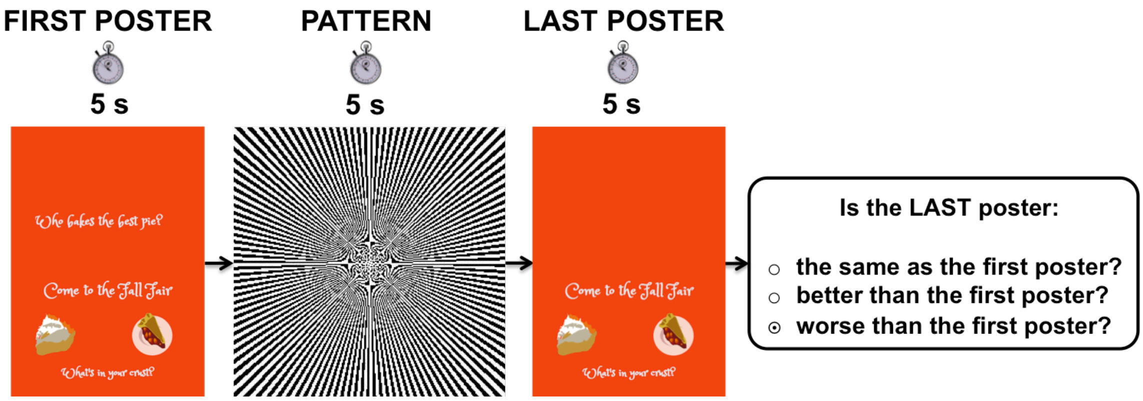

Figure 4. *Recognition post-test: the example task flow of the mini-tutorial*

Post-test Trial Questions

As part of the trial set shown in Figure 5, students were first shown a poster for five seconds (labeled "first poster"), then a pattern for another five seconds (labeled "pattern"), and, lastly, a variation of the first poster for five seconds (labeled "last poster"). Then, students were asked to (1) judge whether the quality of the last poster in the trial set was the same, better, or worse than the quality of the first poster in this trial set, (2) rate the confidence in their answer on a scale of 1 (low confidence) to 5 (high confidence), and, finally, (3) provide a written explanation of their decision.

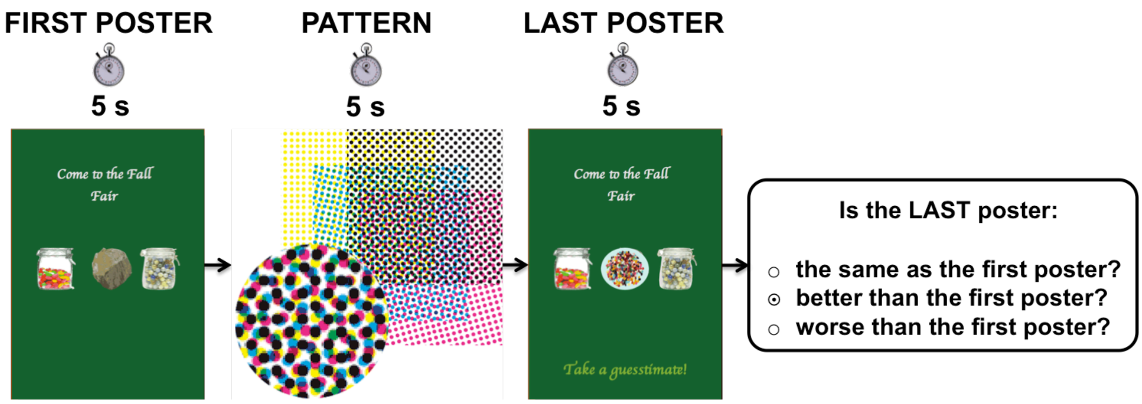

Figure 5. *Recognition post-test: the trial questions of the mini-tutorial. The students were prompted to decide whether the last poster was the same, better, or worse than the first poster, as well as to rate their confidence in their answer, and to write down a justification for their decision.*

Post-test Main Questions

After answering the tutorial and the trial set questions, and getting familiar with the task, students were presented with four sets of images and questions, illustrated in Figure 3. The post-test targets the following four design rules: the text should not be on the poster's edge, the poster should include at least an image, the images should be relevant to the poster's theme, and the text-background color contrast should be high. The post-test evaluates students on three dimensions: their ability to (1) judge posters (i.e., to decide whether the last poster was the same, better, or worse compared to the first poster), (2) rate their confidence in their poster-judging decisions, and (3) justify their poster appraisal decision using graphic design principles.

Internal Learning Choices

In Study 1, students were required to create only two posters, in contrast to the three posters required in the other two studies. Thus, Critical Feedback ranged from zero to six, while Revision ranged from zero to two.

Internal Learning Outcomes

Log file data of students' actions were collected. The post-test constituted of the learning measure, *Poster Ranking*, a dimension of the **Recognition** category. As different posttests were used across the studies, the post-tests aim to test the robustness of the relation between choices and learning outcomes, regardless of the particularities of the learning measures, to demonstrate the validity of the measures included across the studies.

Do Learning Choices Correlate with Internal Learning Outcomes?

Correlations, regressions, and partial correlations were conducted to understand the relations between the choice and learning measures. Results of a repeated-measures analysis of variance, conducted to assess whether there was any improvement in students' poster quality from the first round to the last, revealed that performance (as measured by *Poster Quality*) improved across game rounds: Round 1 = 10.02, Round 2 = 12.75; Wilks' Lambda = .79, $F(1,51) = 13.71$, $p < .01$. Thus, performance can also be considered to be a learning measure, as students improved their knowledge of graphic design principles significantly from the first poster to the last poster. The correlations among the learning choices and internal learning outcomes for Posterlet players are shown in Table 1. In this study, *Poster Ranking* measures students' post-test learning. Results show that *Critical Feedback* correlates with *Revision* and with both internal learning outcomes (*Poster Quality* and

Poster Ranking). Also, *Revision* moderately correlates with poster performance (*Poster Quality*), but not with the post-test (*Poster Ranking*). Finally, *Poster Quality* correlates with the post-test (*Poster Ranking*), providing convergent validity for the learning measures.

Table 1. Study 1 correlations between choices and internal learning outcomes for Posterlet players

Measures	Revision n = 52	Poster Quality n = 52	Poster Ranking n = 52
Critical Feedback	**.66****	**.32***	**.31***
Revision	--	**.32***	.21
Poster Quality		--	**.34***

Note: $^{**}p < .01$, $^{*}p < .05$

Findings revealed that *Critical Feedback* and *Revision* were strongly correlated. To investigate their uniqueness as predictors of learning, stepwise linear regressions were conducted using *Poster Quality* and *Poster Ranking* as separate, dependent variables. For *Poster Quality* performance, the model comprising *Critical Feedback* and *Revision* is significant [F(2,49) = 3.49, p = .04], accounting for 12.5% of the variance in *Poster Quality* (Adjusted R^2=.09), but choices are not significant predictors (*Critical Feedback*: β = .19, p = .28 and Revision: β = .19, p = .28). A model comprising just *Critical Feedback* significantly predicts *Poster Quality* [F(1,50) = 5.75, p = .02], accounting for 10.3% of *Poster Quality* (Adjusted R^2 = .08), *Critical Feedback*: β = .32, p = .02. A model comprising just *Revision* does not significantly predict *Poster Quality*. For *Poster Ranking*, the choice model is not significant, but when only *Critical Feedback* is entered in the regression, the model consisting of *Critical Feedback* significantly predicts the post-test [F(1,50) = 5.24, p = .03], accounting for 9.5% of the variance in post-test (Adjusted R^2 = .08). Thus, *Critical Feedback* is a unique predictor of both learning outcomes in Study 1.

The relation between learning choices and internal learning outcomes was explored through a series of partial correlations. The partial correlation between *Poster Quality* and *Revision*, controlling for *Critical Feedback*, was not significant (r = .15, p = .28, n = 49). The zero-order correlation showed that there was a statistically significant moderate correlation between *Poster Quality* and *Revision* (r = .32, p < .05, n = 50), indicating that *Critical Feedback* mediated the relation between *Poster Quality* and *Revision*. The partial correlation between *Poster Quality* and *Critical Feedback*, controlling for Revision, was not significant (r = .15, p = .28, n = 49). The zero-order correlation showed that there was a statistically significant moderate correlation between *Poster Quality* and *Critical Feedback* ($r = .32$, $p < .05$, n = 50), indicating that *Revision* mediated the relation between *Poster Quality* and *Critical Feedback.* The partial correlation between *Post-test* and *Critical Feedback*, controlling for *Revision*, was not significant ($r = .23$, $p = .11$, n = 49). The zero-order correlation showed that there was a statistically significant moderate correlation between *Post-test* and *Critical Feedback* ($r = .31$, $p < .05$, n = 50), indicating that *Revision* mediated the relation between *Post-test* and *Critical Feedback.* The partial correlation between *Post-test* and *Revision* controlling for *Critical Feedback*, was not computed because no correlation was found between *Post-test* and *Revision*. Overall, these analyses reveal that *Critical Feedback* predicts and explains both *Poster Quality* and *Poster Ranking* more than *Revision* does. *Revision*, which correlates with *Critical Feedback*, does not predict the internal learning outcomes.

Do Treatment Students Outperform Control Students on the Internal Learning Outcomes?

The study investigated whether playing Posterlet improved students' perception of graphic design principles. An independent-samples *t*-test analysis compared post-test learning between students in the treatment and control condition. On *Poster Ranking*, treatment students (M=1.67, SD=1.06, n=52) outperformed control students (M=1.27, SD=.71, n=41) [$t(88.76)=-2.20$, $p=.03$], as shown in Figure 6.

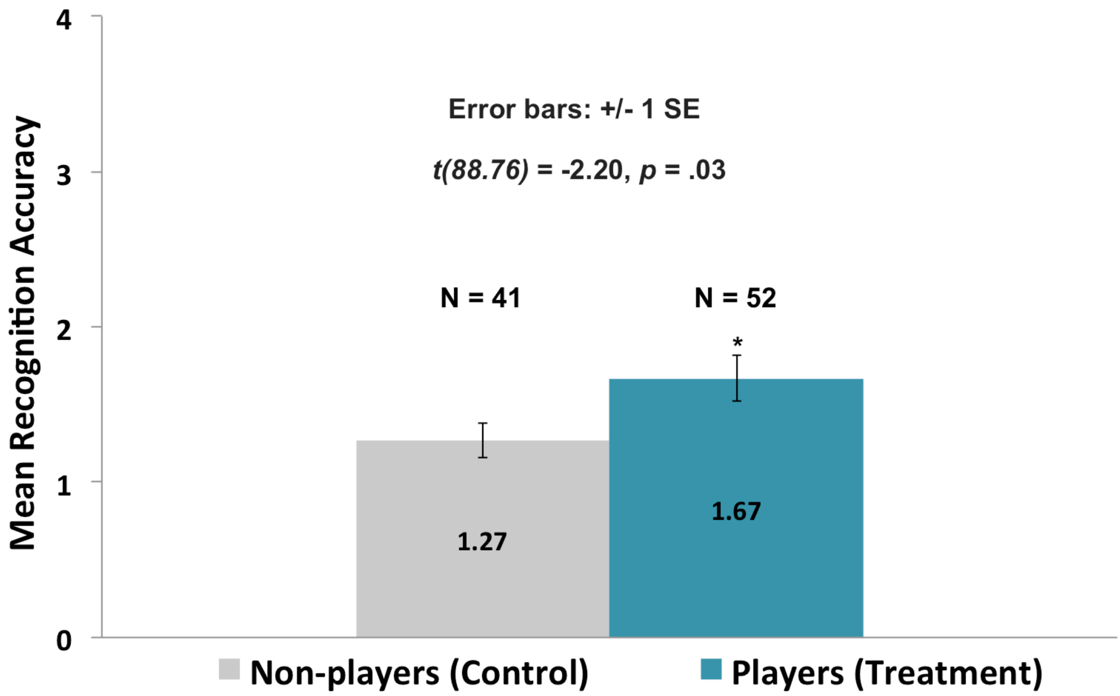

Figure 6. Posterlet players outperform non-players on the post-test

On all questions, except for *Text not on Edge*, treatment students performed above chance (i.e., their scores are greater than -0.33), as illustrated in Figure 7. The value of the scores at chance is -0.33, because every one of the three same/better/worse answers is equally likely with probability 1/3. Each answer item is scored with 1 for the correct answer and -1 for each of the two incorrect answers. The expected value of an answer (i.e., the value at chance) is 1*1/3 + 2*(-1)*1/3 = -0.33.

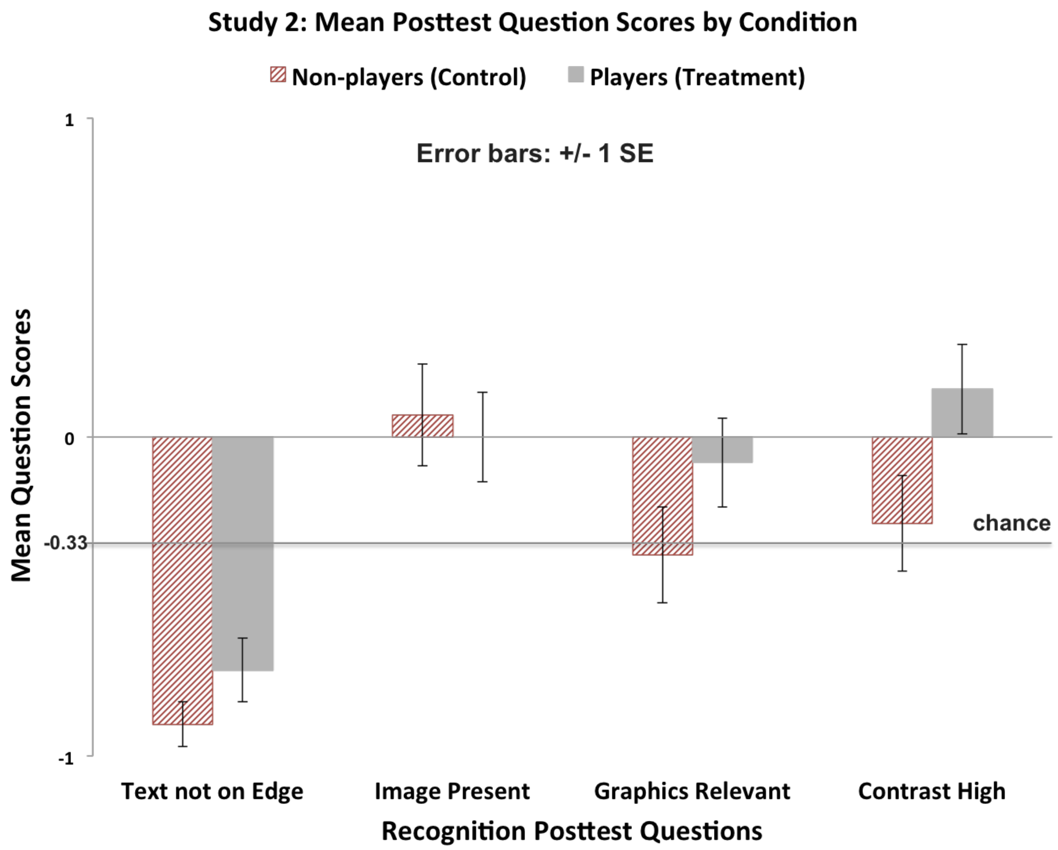

Figure 7. Study 1 post-test scores by condition

Discussion

Results show that students do learn from Posterlet and that it is not the case that they already had knowledge of graphic design principles regardless of the choices they have made in the game, because treatment students outperformed control students on both dimensions of the *Recognition* category of measures. Specifically, treatment students learned to better judge posters and showed increased perception of the graphic design principles. Thus, Posterlet not only helped students learn but, quite possibly, it improved their perception of graphic design principles. This result shows that learning occurred during the assessment and that the assessment did not merely measure problem-solving skills, which constitutes an important distinction.

The study has also found a positive association between seeking critical feedback and both internal learning outcomes, although students played a version of the Posterlet game in which they only designed two posters. This indicates that critical feedback seeking is associated with students' learning outcomes. Findings also show that students' choices to seek critical feedback and to revise were correlated. It may be that students who choose critical feedback also tend to revise their posters and that students who are used to revising more also tend to seek critical feedback more often. However, although learning choices correlate strongly with each other, only seeking critical feedback predicts both internal learning outcomes. Also, findings revealed that performance correlates with post-test learning, providing internal, convergent validity for our learning measures.

Study 2

Participants, Procedure, Design, and Measures

Participants are n = 109 community-college students (63 females, 45 males, and 1 not reported) from California, aged 15-52, $M_{age} = 22.34$ years (SD = 5.67), from the same college as the participants in Study 1. They were recruited as part of a student research subject pool and they participated in exchange for community college course credit. Students provided online consent to participate in the study and they played Posterlet individually (M = 13 minutes), designing three posters, followed by an individual online post-test (M = 3 minutes).

Internal Learning Choices

In this study, students have three opportunities to choose critical feedback on each of the three posters they design. Thus, *Critical Feedback* ranges from 0 (students did not choose any critical feedback) to 9 (students chose only critical feedback). *Revision* ranges from 0 (the student did not revise any poster) to 3 (the student revised all three posters).

Internal Learning Outcomes

The same type of log file data and measures as in Study 1 were collected but a different learning measure was employed. *Poster Quality* measures in-game poster performance, described earlier in the manuscript. *Post-test* measures students' learning of the graphic design principles, which was computed by adding the normalized *Z-scores* of the four post-test questions in the **Critique** (*Common Mistakes* and *Written Feedback*) and **Principle Selection** (*Good Features* and *Bad Features*) categories illustrated in Figure 2.

Do Learning Choices Correlate with Internal Learning Outcomes?

A repeated-measures analysis of variance was conducted to assess any within-student differences in performance outcomes across the study at three time points: first poster, second poster, and third poster. Students' average performance (*Poster Quality*) improved across game rounds, as revealed by a repeated-measures analysis of variance: Round 1 = 9.78, Round 2 = 11.55, Round 3 = 12.10; $F(2,107) = 13.53$, $p < .001$, Wilks' Lambda = .80, partial eta squared = .20. This analysis investigated whether students improved their learning throughout the game, from one round to the next. Results suggest that students improved their learning of graphic design principles from one poster to another, so *Poster Quality*, which expresses the extent to which students use graphic design principles correctly on the posters in the Posterlet game, can also be considered to be a learning measure. Pearson correlations were conducted between the learning choices (seeking critical feedback and revising) and internal learning outcomes (poster performance measured by the game and learning of the graphic design principles measured by the post-test), as shown in Table 2.

Table 2. Study 2 correlations between learning choices (critical feedback and revision) and internal learning outcomes

Measures	**Revision n = 109**	**Poster Quality n = 109**	**Post-test n = 103**
Critical Feedback	**.53*****	**.21***	**.20***
Revision	--	**.37*****	**.33****
Poster Quality		--	**.44*****

Note: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

A test of linearity revealed that the relations between any two of these measures were indeed linear. Both choices (*Critical Feedback* and *Revision*) correlated moderately with both internal learning outcomes (*Poster Quality*, *Post-test*) and strongly with each other. Lastly, *Poster Quality* correlated strongly with the post-test that evaluated graphic design principles, providing convergent validity of the learning measures included in this study. Results revealed that *Critical Feedback* correlated with *Revision*. Therefore, the next analysis investigated whether these choices are individual predictors of learning. Stepwise linear regression analyses were conducted using *Poster Quality* and *Post-test* as separate, dependent variables. For *Poster Quality*, the model composed of the two choices is significant, but only *Revision* enters as a significant predictor [$\beta = .36$, $F(2,106) = 8.55$, $p < .001$], accounting for 13.9% of the variance in performance (Adjusted $R^2 = .12$). For the *Post-test*, the model composed of the two choices is significant, but *Revision* is the significant predictor [β = .31, F(2,100) = 6.02, *p* = .003], accounting for 11% of the variance in learning (Adjusted R^2 = .09). Thus, although both choices taken together predict students' learning outcomes, *Revision* is a unique predictor of both learning measures, while *Critical Feedback* is not.

The relation between learning choices and internal learning outcomes was explored through a series of partial correlation analyses. The partial correlation between *Poster Quality* and *Revision*, controlling for *Critical Feedback*, was moderate (*r* = .31, *p* < .01, n = 106). The zero-order correlation showed that there was a statistically significant moderate correlation between *Poster Quality* and *Revision* (*r* = .37, *p* < .001, n = 107), indicating that *Critical Feedback* had very little influence in controlling for the relation between *Poster Quality* and *Revision*. The partial correlation between *Poster Quality* and *Critical Feedback*, controlling for Revision, was not significant (*r* = .02, *p* = .84, n = 106). The zero-order correlation showed that there was a statistically significant moderate correlation between *Poster Quality* and *Critical Feedback* (*r* = .21, *p* < .05, n = 107), indicating that *Revision* mediates the relation between *Poster Quality* and *Critical Feedback*. The partial correlation between *Post-test* and *Revision*, controlling for *Critical Feedback*, was moderate (*r* = .27, *p* < .01, n = 100). The zero-order correlation showed that there was a statistically significant moderate correlation between *Post-test* and *Revision* (*r* = .33, *p* < .01, n = 101), indicating that *Critical Feedback* had little influence in controlling for the relation between *Post-test* and *Revision*. The partial correlation between *Post-test* and *Critical Feedback*, controlling for *Revision*, was not significant (*r* = .03, *p* = .74, n = 100). The zero-order correlation showed that there was a statistically significant moderate correlation between *Post-test* and *Critical Feedback* (*r* = .19, *p* < .05, n = 101), indicating that *Revision* mediates the relation between *Post-test* and *Critical Feedback*.

Discussion

The same correlation pattern as in Study 1 was found between seeking critical feedback and internal learning outcomes, although the post-test was replaced and students played another version of the Posterlet game in which they only designed a different number of posters. Specifically, in this study, students seeking critical feedback performed better on both internal learning outcomes and they also revised more. Students who chose to revise also performed better on both internal learning outcomes. This indicates that learning choices are associated with students' learning outcomes, regardless of the test used to measure these learning outcomes.

Although learning choices, *Critical Feedback* and *Revision*, correlate strongly with each other, only choosing to revise predicts both learning outcomes in Study 2, while seeking critical feedback predicts both internal learning outcomes in Study 1. Thus, the analyses across the two studies reveal that *Revision* explains both learning outcomes more than *Critical Feedback* does, accounting for a little more of the variance in learning outcomes.

This may be due to different samples between Study 1 and Study 2, in terms of both the age of the participants and the sample size. This may also be due to the different learning outcome measures of the two studies. As Study 2 is a correlational study, it is not clear whether students who revise more are drawn to choosing more critical feedback that would inform their revision process or that students who choose more critical feedback are also more inclined to revise their work. In this study, revising seems more important for learning than seeking critical feedback, being a strong predictor of both internal learning outcomes. It is likely that revising constitutes one pathway from seeking critical feedback to performance, as it provides one mechanism through which students can act on the information they read in the feedback messages.

A crucial difference between Study 1 and Study 2 is the limited opportunity for the students to learn from feedback in only two rounds of the game in Study 1. The first round of the game is always exploratory, students getting used to the mechanics of the game. A third round could have yielded similar results to those obtained in Study 2, as students would have consolidated their learning choice strategy only during the second and third round of the game. Another limitation of both Study 1 and Study 2 is the nature of the participant sample. Students participated in both studies for credit, thus, in some cases, their motivation could have been to complete the task as efficiently as possible, and, consequently to revise less and spend less time designing posters.

In a future study, the revising behavior of students who choose critical feedback will be compared with that of students who are assigned feedback. Such a study will serve to untangle the relation between the two choices (critical feedback seeking and revision). As well, it will clarify the individual relation of choices with the learning and performance outcomes. Finally, results show that students who performed better on the posters also performed better on the post-test, which supports the internal, convergent validity of our in-game learning outcomes. Overall, the findings of Study 1 and Study 2 suggest that learning outcomes can be predicted by students' learning choices.

Study 3

Participants, Procedure, and Design

Participants were n = 97 sixth-grade middle-school students from California (46 females, 51 males), aged 11-12 years. Individually, students designed three posters in the Posterlet game and filled an online post-test. Of all the participants, n = 80 provided both assent and parental consent and completed Posterlet (M = 14 minutes) and, of these, only n = 56 students completed the post-test (M = 7 minutes), due to time constraints.

Post-test Instrument and Measures

Students completed the post-test individually. The task was explained in the introductory paragraph of the post-test. Then, just like in Study 1, they were presented with a trial image set, followed by the main post-test question sets. Additionally, students were asked to answer the two checkbox questions that comprised the **Principle Selection** post-test (*Good Features* and *Bad Features*).

Internal Learning Choices

The same choice and performance measures described in Study 1 and Study 2 were used in this study.

Internal Learning Outcomes

Here, *Post-test* combines the measures of *Poster Ranking* from Study 1 and the **Principle Selection** category from Study 2, described earlier, representing the sum of their normalized *Z*-scores. An alternative learning measure was used to test the consistency of the Study 1 and Study 2 results pertaining to the relation between choices and learning.

External Learning Outcomes

Academic Achievement measures included English Language Arts (*ELA*), Mathematics (*Math*), and Science (*Science*) standardized test achievement.

Do Learning Choices Correlate with Internal Learning Outcomes?

A repeated-measures analysis of variance was conducted to assess potential performance increases from the first to the last game round for each student. Results showed that performance (*Poster Quality*) can be considered to be a learning measure, as it improved significantly from the first poster to the last: Round 1 = 9.76, Round 2 = 11.40, Round 3 = 11.35; Wilks' Lambda = .87, $F(2,78) = 6.01$, $p = .004$. Pearson correlation analyses were conducted to explore the relations among the variables of interest: choices and internal learning outcomes. Results revealed that *Critical Feedback* correlated with *Revision*, as shown in Table 3. Findings showed that only *Revision* correlated with the internal learning outcomes.

Table 3. Study 3 correlations between critical feedback, revision, and internal learning outcomes

Measures	Revision n = 80	Poster Quality n = 80	Post-test n = 56
Critical Feedback	**.54*****	.13	.22
Revision	--	**.23***	**.35****
Poster Quality		--	.02

Note: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Stepwise linear regression analyses were conducted using the learning choices to predict the internal learning outcomes. Similar to Study 2, *Revision* enters as a predictor in the regression for both internal learning outcomes: $\beta = .23$, $F(1,78) = 4.19$, $p = .04$, accounting for 5.1% of the variance in *Poster Quality* (Adjusted $R^2 = .04$) and, similarly, $\beta = .35$, $F(1,54) = 7.65$, $p = .008$, accounting for 12.4% of the variance in *Post-test* (Adjusted $R^2 = .11$). On all questions (except for *Text not on Edge*), students performed above chance (i.e., -0.33), as illustrated in Figure 8.

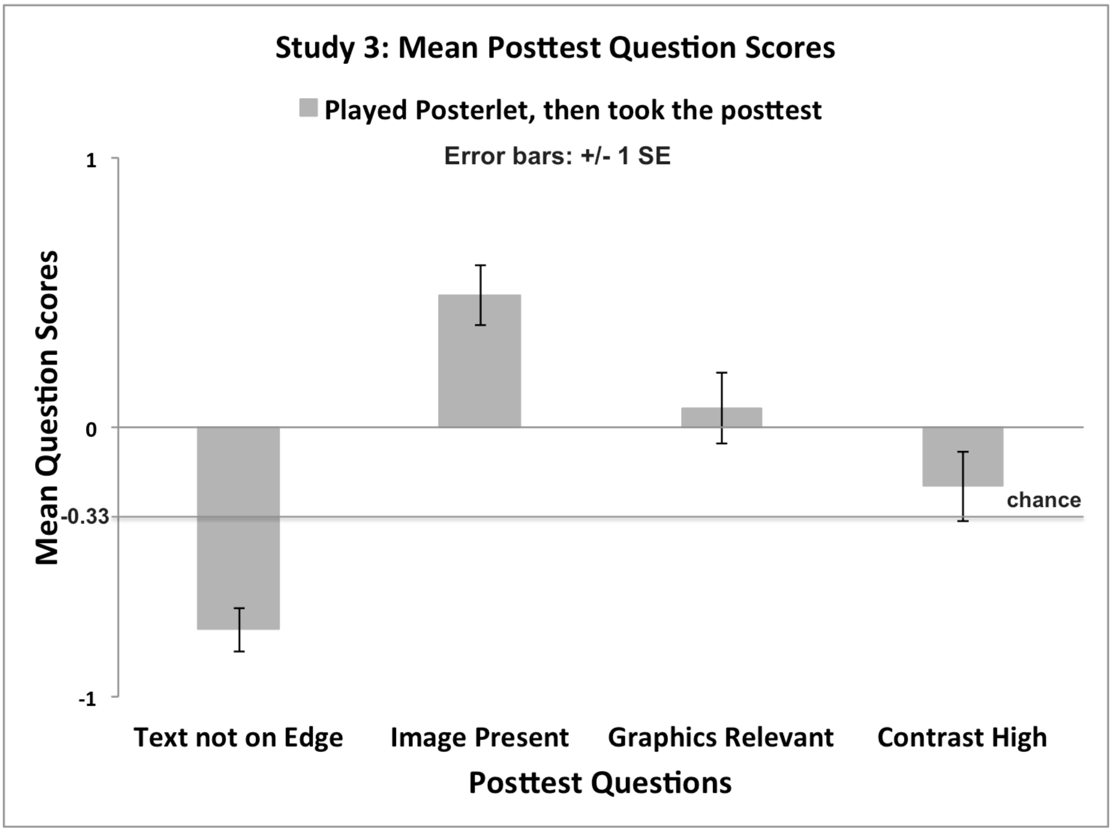

Figure 8. Study 3 post-test scores on each of the four questions

Do Learning Choices Correlate with External Learning Outcomes?

The correlations between internal and external learning outcomes are presented in Table 4.

Table 4. Study 3 correlations between critical feedback, revision, post-test, and outside assessments

Measures	ELA n = 71	Science n = 69	Math n = 42
Critical Feedback	**.39****	**.26***	.30 (p = .05)
Revision	**.30***	.18	**.35***
Post-test	**.42****	**.41****	.30

Note: $^{**}p < .01$, $^{*}p < .05$

Achievement scores were available for many of the students included, although samples varied depending on the achievement scores available. Results indicate stable correlations between learning choices and *ELA* scores. Seeking critical feedback correlates with *Science* proficiency, while revising correlates with *Math*. Thus, internal learning choices correlate with external learning outcomes (i.e., in the school). Also, learning on the post-test correlates with *ELA* and *Science* scores, which supports the external validity of the measures included in this study.

Discussion

Students performed above chance on all post-test questions but the first. Hence, this study provides evidence that students learned design principles from playing Posterlet. The first post-test question was challenging for most students, because it targeted a subtle design rule, *Text not on the Edge*, that not all students encountered as feedback when they played Posterlet. It is also possible that students did not find this particular design rule to hinder their poster's appearance as other more obvious rules (e.g., low color contrast). Results show that learning choices strongly correlated with each other, but only revision predicted internal learning outcomes. In the future, more tuning of the post-test is necessary to ensure that it measures learning of the design principles according only to the design principles that students encountered in their feedback; currently, the post-test samples 4 of the 21 rules, so its quality can be further improved. Both learning choices exercised in Posterlet predicted in-school learning (*ELA*). Additionally, seeking critical feedback predicted *Science* and revising predicted *Math*. Importantly, post-test learning (*Post-test*) predicted in-school learning (*ELA* and *Science*). Moreover, both learning choices and outcomes correlated with ELA, which is possibly the most important academic skill, as it supports reading and comprehension in all other disciplines, including Science and Math learning, and it predicts future reading abilities as well (O'Shaughnessy, Lane, Gresham, & Beebe-Frankenberger, 2003).

This study is of particular importance, as it relates learning choices, performance, and learning measured in our controlled assessment environment to students' learning outside of our assessment, in a real-world setting (e.g., students' school a year before our current study). As relations between in-game learning choices and learning performance with out-of-game learning persisted, this series of studies constitutes a step forward in relating feedback and learning between the two environments. Notably, the degree to which middle-school children learn from critical feedback they choose can be predicted by their previous year's standardized scores. Students' learning performance in school a year before predicts their learning choices and learning in Posterlet, suggesting that students carry their learning behaviors over time from one environment to another. Thus, interventions that can nudge students to choose more critical than confirmatory feedback have the potential to improve students' learning both in controlled environments and in school. In the future, students' Posterlet learning choices and outcomes will be compared with their in-school performance a year later. For instance, researchers sampled 228 students, aged 8 to 25 years, and found that feedback on their learning performance predicted reading and mathematics performance 2 years later (Peters, Van der Meulen, Zanolie, & Crone, 2017). In contrast, finding predictors (e.g., ELA one year earlier) for independent learning strategies such as feedback seeking and revising can help educators nudge students on a trajectory of lifelong learning.

General Discussion and Implications

This research described Posterlet, an instance of a choice-based assessment developed as a measuring and tracking tool for choice behaviors that enables theoretical and empirical explorations of choice behavior measurement in game-based digital environments. The three empirical studies described in this manuscript hypothesized that the choices to seek critical feedback and to revise would predict students' learning outcomes. Across the three studies, findings revealed that the learning choices in the assessment environment correlated with each other and with both learning outcomes, correlations between choices and learning outcomes remaining in the same range (.2 to .4) for bigger samples. These results do not suggest that students who prefer confirmatory feedback did not learn from that feedback, but rather that they learned less than their counterparts who sought more critical feedback.

Results showed through internal validation that choices such as seeking critical feedback and revising are beneficial in predicting learning, although more research is needed to clarify the contribution of each learning choice to the learning outcomes. External learning outcomes (i.e., measures of in-school achievement) were available for the children in the sample. Results revealed through external validation that in-game choices and post-test learning correlated with external outcomes, indicating that choices can predict out-of-game learning as well as in-game learning.

A limitation of the work so far is that, while findings show that choices correlate with learning outcomes, revision predicts internal learning outcomes (poster performance and post-test learning) in Study 2 (college students) and Study 3 (middle-school students), while critical feedback predicts internal learning outcomes in Study 1 (college students). For college students (Studies 1 and

2), seeking critical feedback correlated with the internal learning outcomes (poster performance and post-test learning). The internal learning outcomes also correlated with each other, supporting the internal, convergent validity of the learning measures. However, this result did not generalize to middle-school students, likely due to developmental differences (Peters, Braams, Raijmakers, Koolschijn, & Crone, 2014) or to the different post-test used across the three studies. Instead, for middle-school children (Study 3), the choice to revise correlated with children's internal learning outcomes (poster performance and post-test learning), as well as with their external learning outcomes (standardized achievement scores). Although these studies sampled nearly 300 students, larger-scale studies conducted using Posterlet (Cutumisu, Blair, Chin, & Schwartz, 2015, 2017) showed that both critical feedback and revision choices correlated with learning outcomes and with each other. Lastly, a limitation of this work is the lack of evidence that students would seek critical feedback in social situations, not only from digital characters. More research is needed to shed light on this matter. However, the fact that learning choices can be measured and that these choices lead to improvement in learning outside the game, measured by an independent post-test and by independent academic achievement measures, shows promise that researchers can start investigating the relations between learning choices and learning outcomes in social situations.

Game-based assessments are more entertaining ways than traditional assessments of capturing and assessing learning processes across formal and informal settings over a relatively short time (about 15-30 minutes of gameplay). They are designed to determine whether a learning experience has a measurable effect in changing how students solve challenges in the future, rather than to assess a particular student. Thus, choicelets can be employed as formative assessment tools and they can provide actionable formative validity information that educational practitioners and designers can use to shape their materials or their instruction to improve student learning. This research provided empirical evidence to establish internal validity of choice-based assessments: differences in students' game play choices predicted differences in students' learning outcomes, measured both within and outside of the game assessment environment. Results also provided evidence for external validity of the assessment: students' choices within the game are positively associated with differences in students' learning experiences outside of our assessment environment (e.g., in school). Additionally, Posterlet helped students learn graphic design principles and it differentiated between novice (control) and expert (treatment) students in their abilities to perceive subtleties of design principles. Students who played Posterlet learned to judge posters and to perceive design principles better than control students. Thus, apart from helping students learn, Posterlet may improve students' perception of the world in the context of graphic design principles and poster design. Given these results, this research showed that choices are important, but that the reliability of the assessment is not yet established, in part because the assessment focused more on validating students' choices rather than their learning outcomes.

The implications of this work extend outside the classroom. This research measures individuals' intangible behaviors and their impact on learning. Rather than assessing individuals' knowledge, skills, and abilities at a given moment, one also need to assess their behaviors that could be predictive of their future performance and lifelong learning. For example, in determining whether a candidate is suitable for a job in a fast-moving job market with dynamic requirements, it would be more appropriate to assess the candidate's different behaviors that could predict a range of learning outcomes than to test a narrow knowledge set that may soon become obsolete.

In future studies, the external validation of our measures will be probed by employing academic achievement tests that are independent of the Posterlet intervention (e.g., standardized achievement tests) for all populations included in these studies. Future research directions also include the creation of a catalog of other learning choices that can be measured with such choice-based assessment models. For instance, the use of game-based assessments to evaluate less structured, shorter learning experiences can be explored. In addition to identifying students who could benefit from interventions that could shape their choices and hence their learning, such assessments could also provide formative feedback to educators to help them improve their programs of instruction.

Conclusions

A choice-based assessment game, Posterlet, was employed to track behaviors hypothesized to be important to foster independent learning. Posterlet is a lightweight assessment game that offers players an entertaining 15-30-minute experience and researchers an opportunity to track and measure students' learning processes more naturalistically. By capturing students' free choices in an environment where students have similar experiences and learning opportunities regardless of these choices, this research approaches the goal of measuring how well students are prepared to learn independently in the future. The three studies presented provided empirical evidence that choices to seek critical feedback and to revise predict better learning in the assessment and in school. Thus, this research advocates for embedding learning choices into instructional models in both formal and informal environments to enhance student learning.

Acknowledgements

We would like to thank the students who participated in these studies, as well as their teachers. This work was supported by the NSF Grant EHR-1228831 as well as by the University of Alberta Faculty of Education Support for the Advancement of Scholarship Grant, the Killam Cornerstone Operating Grant RES0043207, and the SSHRC IDG Grant RES0034954. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

Statements on Open Data, Ethics and Conflict of Interest

The data collected for each of the three studies are available together with the manuscript.

This research was carried out under the ethical guidelines and approval from our university ethics review board as well as under the ethical guidelines for educational research of British Educational Research Association (https://www.bera.ac.uk/). The data used in this research was collected subject to the informed consent and assent of the participants.

The authors, Dr. Maria Cutumisu, Dr. Doris B. Chin, and Dr. Daniel L. Schwartz, certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems. ITS 2000. Lecture notes in computer science* (Vol. *1839*, pp. 292-303). Berlin: Springer.

Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *The Journal of General Psychology, 54*(2), 279-299. https://doi.org/10.1080/00221309.1956.9920284

Attali, Y. (2015). Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers & Education, 86*, 260-267. https://doi.org/10.1016/j.compedu.2015.08.011

Black, P. (2015). Formative assessment–an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice, 22*(1), 161-177. https://doi.org/10.1080/0969594X.2014.999643

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*(1), 61-100. https://doi.org/10.3102/0091732X024001061

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219-233. https://doi.org/10.1080/03075070600572132

Chin, D. B., Blair, K. P., Wolf, R. C., Conlin, L. D., Cutumisu, M., Pfaffman, J., & Schwartz, D. L. (2019). Educating and measuring choice: A test of the transfer of design thinking in problem solving and learning. *Journal of the Learning Sciences*. In Press. https ://doi.org/10.1080/10508 406.2019.1570933

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 *Proceedings of the 122nd American Society for Engineering Education (ASEE'15)*, June 14-17. Seattle, WA.

Cutumisu, M. (2018). The informational value of feedback choices for performance and revision in a digital assessment game. *Interactive Technology and Smart Education, 15*(4), 363-380.

Cutumisu, M. (2019, online 2018). The association between critical feedback seeking and performance is moderated by growth mindset in a digital assessment game. *Computers in Human Behavior, 93*, 267–278.

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

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2017, Online first 2016). Assessing whether students seek constructive criticism: The design of an automated feedback system for a graphic design task. *International Journal of Artificial Intelligence in Education (IJAIED)*, *27(3),* 419-447, https://doi.org/10.1007/s40593-016-0137-5, Springer.

Cutumisu, M., & Schwartz, D. L. (2018, Online first 2017). The impact of critical feedback choice on students' revision, performance, learning, and memory. *Computers in Human Behavior, 78*, 351-367, ISSN 0747-5632. https://doi.org/10.1016/j.chb.2017.06.029

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

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. https://doi.org/10.1037/0033-295X.100.3.363

Graesser, A. C. (2017). Reflections on serious games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games. Advances in game-based learning* (pp. 199-212). Cham, Switzerland: Springer.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, England: Routledge, Taylor and Francis Group.

Hattie, J. (2013). *Visible learning for teachers: Maximizing impact on learning* (2nd ed.). London, UK: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64*(4), 349-371. https://doi.org/10.1037/0021-9010.64.4.349

Klehe, U., & Anderson, N. (2007). Working hard and working smart: Motivation and ability during typical and maximum performance. *Journal of Applied Psychology, 92*(4), 978-992. https://doi.org/10.1037/0021-9010.92.4.978

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. https://doi.org/10.1111/1467-8721.ep10772989

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. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97.

Mahfoodh, O. H. A. (2017). "I feel disappointed": EFL university students' emotional responses towards teacher written feedback. *Assessing Writing, 31*, 53-72. https://doi.org/10.1016/j.asw.2016.07.001

Mayer, R. E. (2018). Computer games in education. *Annual review of psychology, 70, 531-549.*

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., . . . John, M. (2014). *Psychometric considerations in game-based assessment.* GlassLab Report: Institute of Play. Retrieved from http://www.instituteofplay.org/work/projects/glasslab-research

Mory, E. H. (2004). Feedback research revisited. *Handbook of Research on Educational Communications and Technology, 2*, 745-783.

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: Perceptions of students and academics. *Assessment & Evaluation in Higher Education, 42*(2), 266-288.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review, 23*, 7-26.

O'Shaughnessy, T. E., Lane, K. L., Gresham, F. M., & Beebe-Frankenberger, M. E. (2003). Children placed at risk for learning and behavioral difficulties: Implementing a school-wide system of early identification and intervention. *Remedial and Special Education, 24*(1), 27-35.

Peters, S., Braams, B. R., Raijmakers, M. E., Koolschijn, P. C. M., & Crone, E. A. (2014). The neural coding of feedback learning across child and adolescent development. *Journal of Cognitive Neuroscience, 26*(8), 1705-1720. https://doi.org/10.1162/jocn_a_00594

Peters, S., Van der Meulen, M., Zanolie, K., & Crone, E. A. (2017). Predicting reading and mathematics from neural activity for feedback learning. *Developmental Psychology, 53*(1), 149-159. https://doi.org/10.1037/dev0000234

Piaget, J. (1964). Quoted by Eleanor Duckworth in "Piaget rediscovered. A report of the conference on cognitive studies and curriculum development". In R. E. Ripple & V. N. Rockcastle (Eds.), *Cognitive studies and curriculum development conference* (1st ed.). Ithaca, NY: Cornell University, ERIC.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267-280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Roozenbeek, J., & van der Linden, S. (2018): The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, 1-11. https://doi.org/10.1080/13669877.2018.1443491

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*(3), 482-486. https://doi.org/10.1037/0021-9010.73.3.482

Sargeant, J., Mcnaughton, E., Mercer, S., Murphy, D., Sullivan, P., & Bruce, D. A. (2011). Providing feedback: Exploring a model (emotion, content, outcomes) for facilitating multisource feedback. *Medical Teacher, 33*(9), 744-749. https://doi.org/10.3109/0142159X.2011.577287

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

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

Schwartz, D. L., Lindgren, R., & Lewis, S. (2009). Constructivism in an age of non-constructivist assessments. In S. Tobias & T. Duffy (Eds.), *Constructivist Instruction* (pp. 46-73). New York: Routledge.

Snow, E., Varner, L., Russell, D., & McNamara, D. (2014). Who's in control? Categorizing nuanced patterns of behaviors within a game-based intelligent tutoring system. *Proceedings of the 7th Conference on Educational Data Mining*, 185-191, July 4-7, London, UK.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology, 25*(2), 207-222. https://doi.org/10.1080/14640747308400340

Starčič, A. I., Lipsmeyer, W. M., & Lin, L. (2019). Using motion capture technologies to provide advanced feedback and scaffolds for learning. In T. Parsons, L. Lin, & D. Cockerham (Eds.), *Mind, brain and technology. Educational communications and technology: Issues and innovations* (pp. 107-121). Cham, Switzerland: Springer.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, UK: Cambridge University Press.

Vygotsky, L. S. (Ed.). (1997). *The collected works of L.S. Vygotsky: Problems of the theory and history of psychology.* New York, NY: Springer, US, Science & Business Media.

Wouters, P., & van Oostendorp, H. (2017). Overview of instructional techniques to facilitate learning and motivation of serious games. Advances in game-based learning. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 1-16). Cham, Switzerland: Springer.

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education, 27*(4), 458-477. https://doi.org/10.2307/749877

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.