Educating and Measuring Choice:
A Test of the Transfer of Design Thinking in Problem Solving and Learning

Doris B. Chin[1,2], Kristen P. Blair[1], Rachel C. Wolf, Luke D. Conlin[3], Maria Cutumisu[4], Jay Pfaffman, and Daniel L. Schwartz

Graduate School of Education, Stanford University

[1] Joint first authors

[2] Corresponding author

[3] Currently at *Department of Chemistry and Physics, Salem State University*

[4] Currently at *Department of Educational Psychology, Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton*

(In press) Journal of the Learning Sciences

This material is based upon work supported by the National Science Foundation (Grants #1228831, #1337414) and the Gordon and Betty Moore Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

Correspondence concerning this article should be addressed to:

Doris B. Chin, Graduate School of Education, Stanford University, 450 Serra Mall, Building 160, Room 430, Stanford, CA 94305, USA

Email: dbchin@stanford.edu

URL: aaalab.stanford.edu

Abstract

Educators aim to equip students with learning strategies they can apply when approaching new problems on their own. Teaching design-thinking strategies may support this goal. A first test would show that the strategies are good for learning and that students spontaneously transfer them beyond classroom instruction. To examine this, we introduce choice-based assessments (CBAs). CBAs measure how people learn when there is minimal guidance and they must make independent, learning-relevant decisions. Sixth-grade students completed multiple design activities that either emphasized seeking constructive criticism or exploring a space of alternatives. Afterwards, they completed the CBAs, which measure strategy transfer. Results show that lower-achieving students benefitted most from instruction, exhibiting a relative increase in their use of design-thinking strategies. Additionally, strategy choices correlated with prior achievement measures *and* appeared to mediate performance in and learning from the CBAs. The choices to use the two strategies themselves were uncorrelated, indicating that they are not subsets of a larger construct, such as growth mindset. In sum, CBAs enabled a double demonstration: design-thinking strategies may improve learning and problem solving; and, design-thinking instruction may improve the likelihood of lower-achieving students choosing to use effective strategies in novel settings that require new learning.

Key words: assessment, learning technology, learning strategies, learning-to-learn, design thinking, self-regulated learning, transfer, feedback, 21st-century skills

Educating and Measuring Choice:
A Test of the Transfer of Design Thinking in Problem Solving and Learning

Students do not leave school fully formed, masters of all they could possibly need to know. People change jobs, and the world changes around them. Ideally, students can continue problem solving, learning, and adapting to face new challenges. Students need preparation for future learning. Contexts of future learning may not mimic the classroom, and people will need to learn without the strict guidance of a parent or teacher. When they leave school, students will need to make choices about what, when, and how to learn, often in a context that does not present textbook-organized material, clear problem statements, and highlighted relevant concepts.

Educators have put forth many proposals for the competencies and dispositions that students need for future success. Lists of "21st-Century" skills and dispositions include broad constructs such as critical thinking, creativity, and collaboration (Trilling & Fadel, 2009; Rotherham & Willingham, 2009). They are an attempt to define what students need for an unknown and dynamic future. As school districts adopt "soft skills" into their mission statements and curriculum, one question is whether students will use them in the future when not explicitly told to do so. A second question is whether these skills actually improve performance and learning.

To address these issues, learning scientists need the ability to make direct assessments of the key processes and behaviors that they hypothesize may support learning. Researchers in various fields, from socio-cognitive traditions to metacognition, have lamented the lack of measurement tools that go beyond surveys or distal achievement post-tests that focus on correct knowledge (see Schwartz, Cheng, Salehi, & Wieman, 2016).

To support research into these questions, we introduce a measurement tool called *choice-based assessments* – CBAs (Schwartz & Arena, 2013). CBAs are interactive technologies that present people with a challenge and learning resources, and then track how users go about learning to solve the challenge. Ideally, they capture how students choose to learn when there is neither strict guidance nor strong cues to appropriate behavior. Pintrich and De Groot (1990) captured one key issue with their phrase, “skill and will.” It is not enough for students to know a strategy to solve a problem (the “skill”). They must also choose to use it (the “will”). For instance, students may realize that answering practice questions in math improves learning. Nevertheless, they may skip to the back-of-the-book solutions because it is less effortful.

For life beyond school, educating strategic choice is a critical goal. To know if we are educating choice well, we need the ability to measure it. Here, we demonstrate the use of CBAs in the context of teaching and measuring the benefits of design-thinking strategies for problem solving and learning (Noweski et al, 2012: Razzouk & Shute, 2012).

Learning Strategies and Design Thinking

There are many complementary ways to educate students for future learning. Some examples include improving general cognitive abilities such as executive function (Diamond & Lee, 2011); fostering productive motivational profiles with growth mindset or values affirmation interventions (Dweck, 2006; Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009); strengthening foundational literacies like reading and writing (Spörer, Brunstein, & Kieschke, 2009; Couzijn, 1999); and, helping students develop an understanding of broadly applicable concepts (Schwartz & Martin, 2004).

A different approach comes from scholars who have focused on strategies and heuristics that support independent learning. Much of this literature comes under the banners of “learning to learn” and self-regulated or metacognitive learning strategies (cf. Wirth & Perkins, 2008; Hoskins & Fredricksson, 2008; Zimmerman, 2002; Pintrich & DeGroot, 1990; Ridley et al., 1992). These frameworks include domain-general strategies that apply to many topics of school study, but they are context-specific in that they target school-like tasks. Examples include learning from texts with strategies for summarization (Brown, Campione, & Day, 1981) and self-explanation (Chi, Leeuw, Chiu, & LaVancher, 1994); improving mathematical problem solving with goal setting and self-evaluation (Fuchs et al., 2003); and, learning practices of scientific argumentation (Osborne, Erduran, & Simon, 2004). There has also been recent work on strategies for learning from online environments. For instance, researchers have developed a metacognitive tutor to support better help-seeking behaviors in a computer-based tutoring environment (Roll, Aleven, MacLaren, Koedinger, 2011). Others have created teachable agent systems (learning environments in which students are the tutors for computer characters) to support information structuring and monitoring behaviors (Biswas et al., 2010).

Here, we focus on learning strategies from the burgeoning movement of design thinking in education (Goldman & Kabayadondo, 2017). Design thinking comprises a set of activities that broadly include problem definition, brainstorming, and planning, followed by iterative cycles of creation, testing, and evaluation. Though more typical of life beyond school, many educators embrace design thinking as an antidote to pedagogies that emphasize single correct answers and narrow definitions of academic achievement (Carrol et al., 2010; Koh, Chai, Wong, & Hong, 2015). Design-based activities appear in science, math, engineering, as well as more interdisciplinary, project-based learning (Kolodner et al, 2003; Silk et al, 2010; Sadler, Coyle, & Schwartz, 2000; Bamberger & Cahill, 2013; Conlin, Chin, Blair, Cutumisu, & Schwartz, 2015; Barron et al., 1998). Educators are also anticipating that design-thinking skills and dispositions are likely demands for future careers, where non-routine problems will require new learning in the service of innovation. Design-thinking strategies evolved to improve product design, but the hypothesis is that they may also deliver skills and attitudes that improve abilities to adapt and learn.

There are different suggestions for the proper elements and sequencing of activities in design (Bamberger & Cahill, 2013). Some approaches emphasize empathy in a user-centered process, whereas others emphasize the flare of brainstorming to generate many ideas followed by the focus of prototype production and iterative design cycles to maximize innovation (Brown & Wyatt, 2010). Despite differences, a singular quality unifies design-thinking methods. They provide strategies that protect people from natural tendencies that can inadvertently interfere with effective design. Empathy in design, for example, prevents the natural tendency to over-generalize one’s own experience. Interestingly, scientific methods also have a similar protective quality. The strategies of evidence gathering, replication of experiments, and comparison to statistical thresholds prevent the tendency to accept one’s intuitive beliefs or authority. For both science and design, many methods exist for protecting people from natural tendencies and their associated choices, which can interfere with new learning and discovery.

One tendency, which we focus on here, is the pull towards early closure (Acredolo & Horobin, 1987). Early closure refers to settling on a final solution prematurely. In a classic study, Csikszentmihalyi and Getzels (1970) showed that studio artists who rapidly began painting a still life produced less creative works compared to artists who took their time to consider different possible arrangements first. People have a strong tendency to follow their first good idea, rather than keep an open mind and pursue other alternatives. People may become attached to their ideas or may want to expedite the task and thus not explore the space fully. We focus on two design-thinking strategies intended to block early closure: exploring a space of alternatives and seeking constructive criticism.

Our key empirical questions are whether students will spontaneously transfer the choice to use these two design-thinking strategies after instruction, and whether the strategies support performance and learning (not just product design). To address these empirical questions, we developed CBAs that capture whether students choose to use the strategies when left to their own devices. It is one thing for students to use a productive strategy when there are strong cues and incentives to do so, yet it is quite another for them to use it spontaneously in a new situation.

Two Design-thinking Strategies and Their Associated Choice-Based Assessments

We focus on two design-thinking strategies that help prevent early closure and that we hypothesize are beneficial for learning. The two strategies are seeking constructive criticism instead of only praise and exploring multiple possibilities before settling on a final solution.

These two strategies address different psychological tendencies. Avoiding criticism, constructive or otherwise, is reasonably associated with ego protection (Klugar & Denisi, 1998). In contrast, choosing a first solution, rather than exploring a space of alternatives, is associated with over-commitment to one's ideas and satisficing (Csikszentmihalyi & Getzels, 1970). Strategies that help people avoid these psychological tendencies should benefit problem solving. Moreover, we hypothesize that the strategies may also improve independent learning. Seeking constructive criticism can alert learners to opportunities for improvement. Similarly, exploring multiple possibilities before settling on a solution can help people develop a better sense of the features that contribute to a solution. As described below, CBAs include new content that students can learn (e.g., graphic design principles). This way, we can determine if students make strategic choices in the context of learning and if those choices actually improve problem solving and learning. For the two main CBAs which we describe in the following subsections, we capture three measures: a strategy choice measure, an in-game problem-solving (performance) measure, and a post-test learning measure that evaluates what students learned about the content in the CBA (e.g., graphic design principles). We hypothesize that strategy choices will affect both performance and learning.

We purposely chose to examine two design-thinking strategies that address different psychological aspects of early closure. This allows us to generalize our findings across two instances of design-thinking strategies and two different CBAs. Moreover, the degree of correlation between the use of the two strategies helps determine whether they measure different constructs or whether they tap into the same underlying psychological profile. In the following sections, we describe the two constructs. We also describe the associated assessments in some detail because they do not employ a typical measurement template (e.g., Likert scale, multiple choice) and examples may be informative.

Seeking Constructive Feedback

Feedback is important for learning and motivation (Ammons, 1956; Kluger & DeNisi, 1998; Ilgen, Fisher, & Taylor, 1979; Mory, 2003). Informative negative feedback, which connotes constructive criticism (not punishment), is often more effective for continued learning than positive feedback (Kluger & DeNisi, 1998). Positive feedback may signal sufficient effort toward a task and contribute to early closure. Constructive criticism indicates the need for a change (Hattie & Timperley, 2007). However, criticism also runs the risk of triggering an ego threat that causes retraction rather than learning (Chase, Chin, Oppezzo, & Schwartz, 2009; Kulik & Kulik, 1988). This suggests that students' attitudes towards seeking feedback could have meaningful implications for learning.

Research on feedback originally arose from the behaviorist tradition (Thorndike, 1927). Nearly all feedback studies focus on supervised feedback in which it is up to the teacher, experimenter, or computer to decide when and how to deliver feedback. In feedback research, students rarely have independent control over their feedback, and the feedback arrives without choice (but see Roll, Aleven, McLaren, & Koedinger, 2011; Finkelstein & Fishbach, 2012). Nevertheless, in many situations it may be useful for people to seek feedback. Little is known about the implications of people's feedback choices for their learning. While there is a reason to believe that attitudes towards feedback influence learning, there has been little evidence whether independent choices about feedback are important for learning. Design-thinking methods that focus on embracing constructive criticism could be particularly useful for helping students who might otherwise avoid negative feedback, if those design-thinking strategies succeed in transferring to new situations.

Poster Making: A feedback-seeking assessment. To measure student strategy choices when seeking feedback, we developed the Poster Making-Feedback CBA, which we shorten to Poster-Feedback for the remainder of the article. In Poster-Feedback, players read that they are on the planning committee for their school's Fall Fun Fair (see Figure 1A for the flow of the assessment). Their task is to design posters for three of five activity booths. For each poster, they have a multitude of design options including text and image palettes. The text palette provides a set of phrases (e.g., the time and place of the event, "Come to the Fall Fun Fair!"). The image palette provides a variety of pictures that players can choose from. Players may place items anywhere on the poster and can customize each item (e.g., text font, color, alignment, image size). When their design is complete, players press the "Test Poster" button. The program takes the players to a viewing room where 12 animal characters observe their poster. Players select three of the animals as their "focus group" to gather feedback (Figure 1A, step 3). Critically, for each of the three characters, players must choose to receive *either* positive or negative feedback (Figure 1A, step 4). Players see boxes above each character that say, "I like…" and "I don't like…." and can only choose one box for each character (Figure 1B).

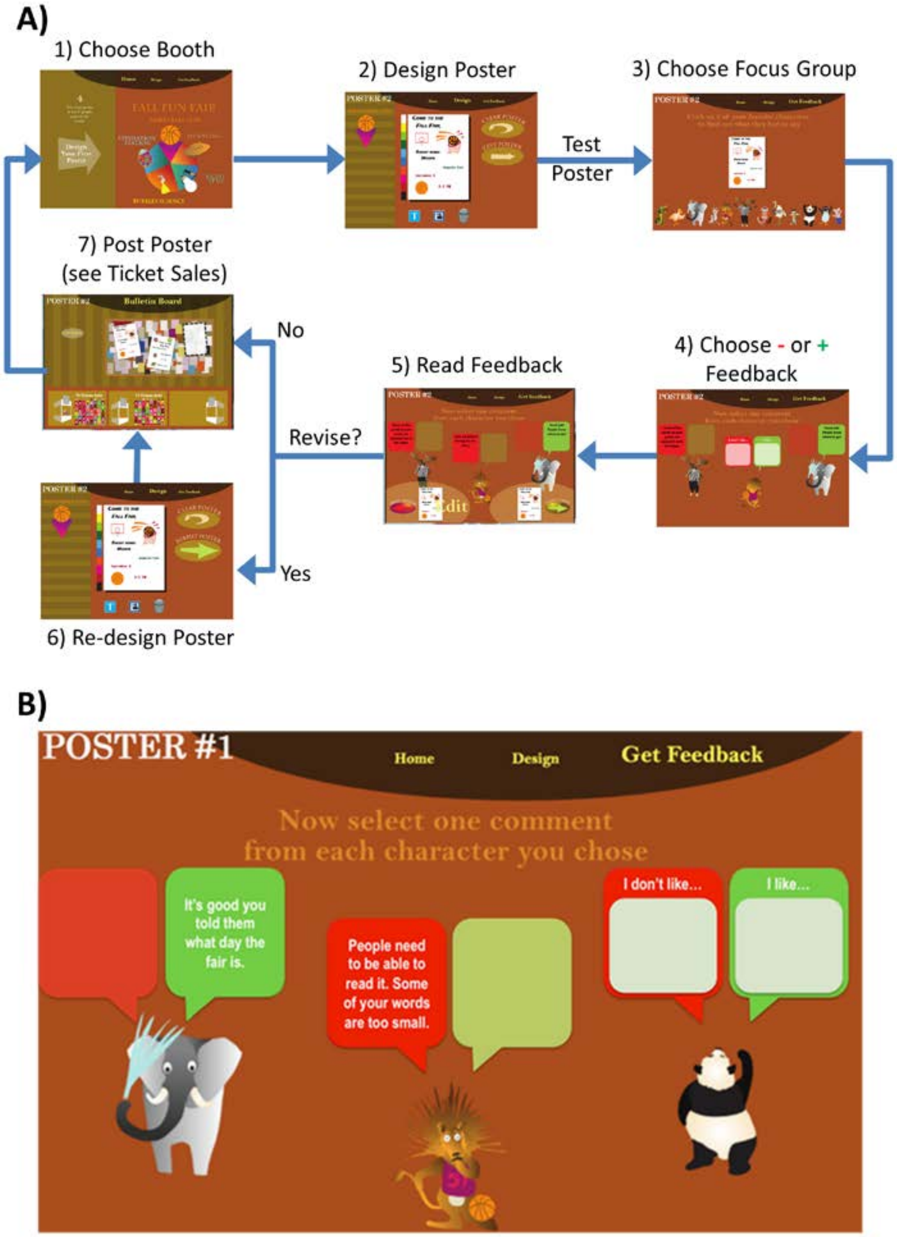
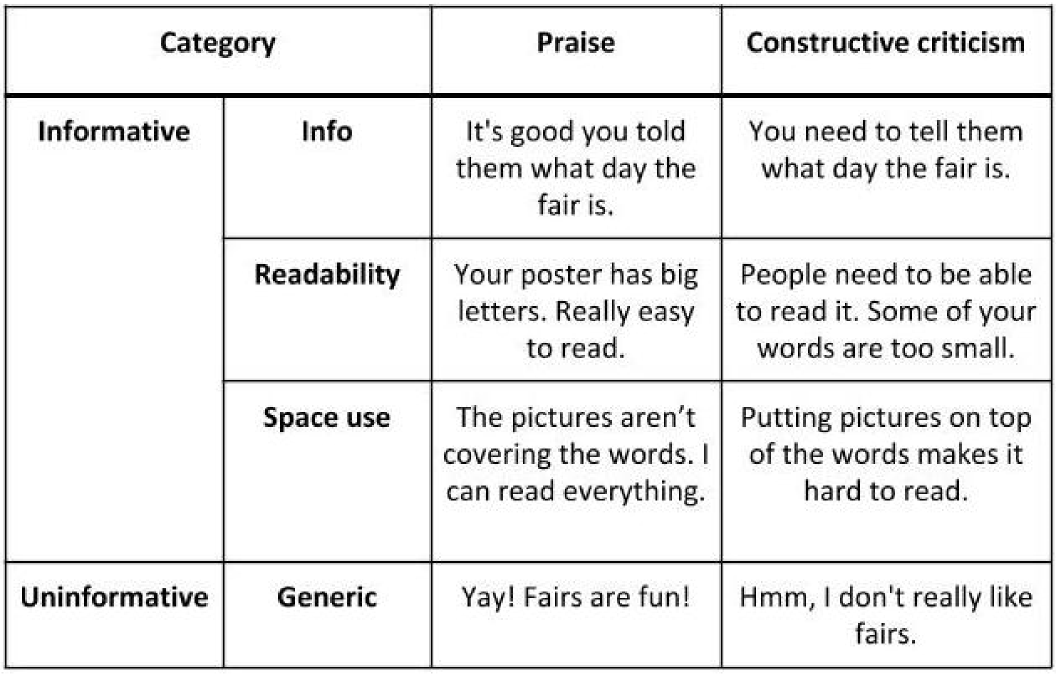

A)
1) Choose Booth
2) Design Poster
Test Poster
3) Choose Focus Group
4) Choose - or + Feedback
5) Read Feedback
Revise?
No
Yes
6) Re-design Poster
7) Post Poster (see Ticket Sales)
B)
POSTER #1
Home
Design
Get Feedback
Now select one comment from each character you chose
It's good you told them what day the fair is.
People need to be able to read it. Some of your words are too small.
I don't like...
I like...

Figure 1. A) Schematic of the flow of the Poster Making – Feedback game. B) Close-up of the Choose Feedback screen. This player chose praise feedback from one character ("I like…"), constructive criticism from another ("I don't like…"), and is in the process of choosing feedback from a third character.

The Poster-Feedback system runs an intelligent engine in the background that evaluates the players posters along 21 pre-specified criteria that check for content (e.g., inclusion of necessary information or appropriateness of image for activity) and general graphic design principles (e.g., proper sizing of text, spacing, color contrast). The system selects from a pool of feedback messages, depending on the evaluation of a poster's features and the players' requests for positive or negative feedback. (The system gives the same feedback regardless of which character is chosen.) We designed the positive and negative feedback to be equally informative. The message pool also includes uninformative generic feedback so the system can respond if it is unable to identify any relevant feedback. Table 1 shows examples of the feedback (for more details of the system, see Cutumisu, Blair, Chin, & Schwartz, 2016).

After reading the feedback (step 5), players choose whether to revise or submit their poster. When players choose to iterate and revise (step 6), they do not receive a second round of feedback on their revision. There is only one round of character feedback and revision for each poster. Upon poster submission (step 7), players receive their poster's score as a number of tickets sold for that booth. The cycle then repeats with players choosing a second and then third booth. Overall, players make 3 posters, leading to a total of 9 feedback choices and 3 revision choices.

Table 1. *Types and examples of feedback in the Poster-Feedback assessment.*



Category		**Praise**	**Constructive criticism**
Informative	**Info**	It's good you told them what day the fair is.	You need to tell them what day the fair is.
	Readability	Your poster has big letters. Really easy to read.	People need to be able to read it. Some of your words are too small.
	Space use	The pictures aren't covering the words. I can read everything.	Putting pictures on top of the words makes it hard to read.
Uninformative	**Generic**	Yay! Fairs are fun!	Hmm, I don't really like fairs.

In a previous research study, several hundred middle-school students in Chicago and New York City played Poster-Feedback (Cutumisu, Blair, Chin, & Schwartz, 2015). The primary question was whether the strategy to select constructive criticism (negative feedback) correlated with better problem solving and learning outcomes. Unlike many assessments that evaluate objectively right and wrong answers (e.g., 2+2 really does equal 4), we needed to develop evidence whether choosing constructive criticism is a relatively good choice for learning. We did this directly by analyzing results from the post-test of graphic design principles appended to the end of the CBA (e.g., in one item, students saw a poster, and they had to check off the good and bad features). We also did this indirectly in post-hoc analysis by correlating the frequency of choosing constructive criticism with standardized achievement test scores.

In this previous study, students who chose more constructive criticism exhibited better problem-solving performance (more improved posters over time in the game) and better learning (performed better on the post-test). Additionally, the frequency that students chose negative feedback showed a consistent positive correlation with standardized test scores. The correlations between feedback choices and achievement in math and English ranged from r = 0.33 to r = 0.41 across the four different standardized tests (New York and Illinois use different standardized tests). In contrast, the frequency that students chose to revise did not exhibit consistent correlations across the tests. Thus, seeking constructive criticism correlated with better learning in the CBA and in school. Inversely, seeking praise negatively correlated with learning and achievement (Cutumisu, Blair, Chin, & Schwartz, 2015). These findings were purely correlational, so a latent trait, such as IQ or growth mindset, may have driven the correlations. Therefore, one purpose of the current research is to determine experimentally whether the choice to seek constructive criticism is malleable and responsive to a design-thinking intervention.

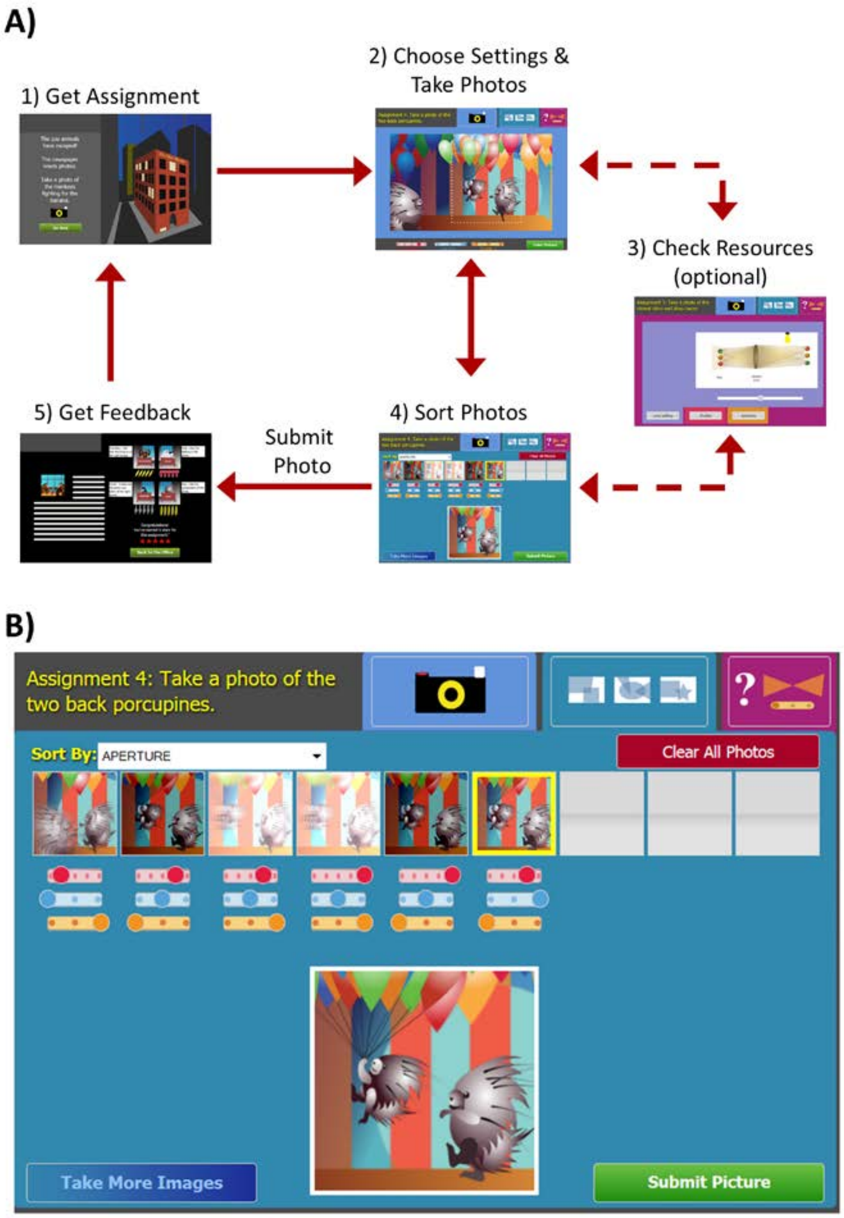
Exploring a Space of Alternatives

As a design-thinking strategy, designing multiple solutions in parallel leads to a fuller exploration of the problem space and better products as compared to refining one's initial idea. Dow, et al. (2010) asked participants to design a web-banner advertisement for an online journal. In the parallel design treatment, participants produced three different designs. They received three pieces of generic feedback, such as "Always pay attention to spacing," and then produced their final design. In the serial design treatment, participants produced a design and received one piece of generic feedback, repeated the design-feedback cycle two more times, and then made their fourth and final design. Experts rated the final advertisements from the parallel treatment more highly than those in the serial treatment. The designs from the parallel treatment also led to more user “clicks” on the advertisements when posted on a popular social media website. The parallel-design technique led the participants to explore more of the design space in each of their attempts, so they had a broader set of options to consider in their final design. The serial-design method let participants slip into the common strategy of refining a first design, so they did not explore as broadly and their products suffered.

Exploring the space should also improve learning because people have the opportunity to discover new properties and possibilities. For example, Dunbar and Klahr (1989) asked children and adult participants to program electric cars using an unfamiliar programming language. Without instruction, only participants who conducted a wide initial search of the parameter space ultimately discovered how a novel function worked (see Bonawitz et al., 2011, for an example with infants).

Photo Taking: An exploration assessment. To assess students’ choices to explore a space of alternatives, we created the Photo Taking-Explore CBA, or Photo-Explore for short. In Photo-Explore, students are junior photographers for the *Oh No! Gazette*. The problem context is that all the zoo animals have escaped and are running amok; players must photograph scenes that capture the mayhem.

Figure 2A shows the game mechanics of the assessment. Players first receive an assignment from the editor to catch specific animals in action (Figure 2A, step 1). Next, the system sends players into the field where the animals are creating chaos in various scenarios (e.g., porcupines meandering in the balloon shop). Players endeavor to get a high-quality action shot as the characters move across the screen. They can adjust the camera’s position and settings (step 2).



Assignment 4: Take a photo of the two back porcupines.

Sort By: APERTURE

Figure 2. A) The flow of gameplay in the Photo Taking - Explore game. B) Close-up of the sort tab for Level 4. The student took six photos and is in the process of examining and rearranging them.

When players click on “Take Picture,” the system briefly displays a photo on the screen before the action continues and players can take more pictures. At any time, players have the option of clicking on the resources tab to look at animated explanations of different camera settings’ functions (step 3). Once they have taken one or more photos, players can access the sorting tab (step 4, and Figure 2B). Here, they can view their photos in closer detail and sort through them. They can also choose to clear their photos, select a photograph to submit to the editor, or to return to the scene and take more pictures. Clicking "Submit Picture" leads to a screen that provides feedback about the photo based on its composition, lighting, and blurriness (step 5). After providing feedback, the system sends players to their next photography assignment.

The CBA includes four task levels. The first level involves composition and focus depth. Students set the composition by positioning the camera so the targets are in view, and they change the focus by adjusting the lens setting slider. (For sliders, students do not see changes in real time as they adjust the slider– they only see the results when they take a picture.) In later levels, students need to adjust the shutter speed and aperture sliders to get the desired photographs. These levels add the challenges of motion blur (from character movement), lighting (over- and under-exposure), and depth of field (getting characters at multiple depths in focus simultaneously). Figure 3 gives an overview of the four levels.

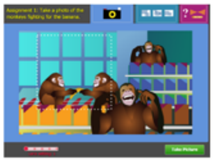
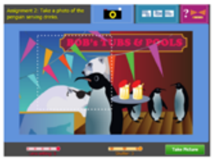
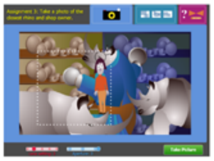
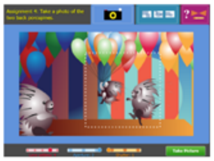
Level	**1**	**2**	**3**	**4**
Assignment: Take a picture of...	Two monkeys in front of the window, fighting for a banana. 	The penguin serving drinks. Penguin is in foreground, moving across the screen. 	The closest rhino and the china shop owner. Rhinos are moving across screen. 	The two porcupines in back. Farthest porcupine is moving across the screen. 
Photographic elements in play				
• Composition	•	•	•	•
• Focus depth	•	•	•	•
• Exposure		•	•	•
• Motion blur		•		•
• Depth of field			•	•
Camera settings to adjust				
• Camera position	•	•	•	•
• Lens setting (5 positions)	•	•	•	•
• Shutter speed (3 settings)		•		•
• Aperture (3 settings)			•	•

Figure 3. Descriptions of the four assignments in the Photo Taking – Explore game with the relevant photographic elements to learn and the available camera settings at each level.

Like the Poster-Feedback CBA, the Photo-Explore CBA affords student choice, as there are many possible paths to a successful photograph. Of interest is whether students choose the strategy of exploring a space of alternatives for taking effective pictures. Specifically, how many unique camera settings do the students explore before settling on the photo to submit? The system measures problem-solving performance by the quality of the photos taken, using an internal engine for evaluating focus, lighting, motion blur, and depth of field. To address learning, as with the poster task, we appended a post-test to determine if choosing to explore a space of possibilities correlates with better learning about photography fundamentals. Students see seven pairs of real photos and must determine what camera settings have changed between the photos in each pair. They also receive four multiple-choice questions that probe their declarative understanding of camera setting effects. Below is a sample question:

> What happens when the shutter speed is slow (check all that apply):
>
> (a) The image gets brighter.
>
> (b) The image gets darker.
>
> (c) Moving objects will be more blurry.
>
> (d) It freezes the movement of objects.
>
> (e) More things are in focus (both near and far away).
>
> (f) The lens needs to be re-focused.

As described, the photo task also includes a resource tab that presents information relevant to learning about the mechanisms by which camera features, such as shutter speed, relate to image outcomes. The resource provides a learning opportunity, but the choice to understand mechanism is orthogonal to our primary construct of interest – exploring a space of alternatives, and we do not discuss it here.

Apex Consumer: A simpler choice-based assessment of exploring the space. Unlike Poster-Feedback, which had been previously vetted, this study was the first deployment of Photo-Explore. Therefore, we included an additional choice-based measure to provide convergent validity that the design-thinking curriculum influenced students’ likelihood of exploring a problem space. We refer to this assessment as Apex Consumer-Explore, or Apex-Explore for short. This measure offers an example of a less elaborate CBA that uses the online survey software *Qualtrics*.

The Apex-Explore CBA is designed for students who have learned about producers and consumers, but not about apex consumers (a predator, such as a shark, that is not eaten by other consumers). Students read that Dan, a hypothetical student, has proposed a new category of consumer in the food chain and their task is to figure out the mystery category. They see a picture of a tiger and read that a tiger belongs to Dan's new category. They then choose between seeing an example of an animal that fits the category, seeing an animal that does not fit the category, or making their guess of Dan's category. Students can see up to five examples total, but once they choose to make their guess, the assessment ends. The key strategy measurement is how many examples students choose to see before guessing; we assume that choosing more examples is an analog of exploring a problem space. Apex-Explore is a secondary source of information about strategy choices and does not include a performance or learning measure.

Comparison to Other Process-Oriented Assessments in Learning Technologies

CBAs share similarities to other learning technologies that collect and analyze process data; they also have some notable differences. One primary difference is that the assessments do not reside within a specific curriculum, and therefore they can compare the effects of different courses of instruction. Many computer-based learning environments, such as intelligent tutoring systems (ITS; Koedinger, Anderson, Hadley, & Mark, 1997), measure learning as students work through the curriculum. The systems can evaluate student knowledge states over time, aiming to determine the best sequence of problems or tutor moves (e.g., intervene with a hint) that will maximize learning from the tutor (Chi, VanLehn, & Litman, 2010). Recently, this work has further begun to consider affect (e.g., boredom and engaged concentration), gaming the system, and off-task behaviors (Fancsali, 2014; Baker, Corbett, & Koedinger, 2004; Baker, Corbett, Roll, & Koedinger, 2008; Roll, Baker, Aleven, & Koedinger, 2014). Their primary purpose is to determine ways to help students learn more from the system. This is of high importance but not ideal for an assessment intended to examine outcomes outside the original context of instruction.

A second difference is that CBAs have *a priori* assessment goals that are close to the surface rather than inferred through extensive data mining. Data mining techniques have been employed to examine students' learning choices in games (Peacock et al., 2013; Lee, Liu, & Popović, 2014; Snow, Allen, Russell, & McNamara, 2014). With the exception of Snow et al. (2014), this research also focuses on predicting future student moves or likelihood of success within the system. For assessment, a challenge of data mining techniques is that the claim connecting behaviors to constructs often requires a complex chain of data transformations that is difficult for stakeholders to interpret (e.g., a clustering algorithm that creates an abstract centroid in a multi-featured space). Therefore, we have collapsed the distance between the raw behaviors and their assessment interpretations, for example, by simply counting how often a student chooses negative (versus positive) feedback.

A third difference is that CBAs are relatively short. In one assessment study, Shute, Ventura, Bauer, and Zapata-Rivera (2009) had an *a priori* assessment goal of measuring student persistence using the physics learning game *Newton's Playground* (Ventura & Shute, 2013). Consistent with their goal of grounding a new approach to assessment, they further correlated the results of their assessment against an existing measure of persistence. This assessment depended on several hours of a student's interactions to collect sufficient data, as the goal was to adapt the game dynamically to address individual student persistence needs. We have opted for assessments of 10-15 minutes. This increases the flexibility of deployment, although we gather less information about any given student.

A Study on the Generalization of Design-Thinking Strategies

Overview

The current experiment evaluated whether it is possible to measure differences in the strategic choices people make in the context of design. Embedded in this evaluation is whether design-thinking strategies transfer beyond the original conditions of learning and whether design-thinking methods are useful for performance and learning. Sixth-grade children completed similar design activities using the same design cycle across curricular units in their math, science, and social studies classes. At select points, the instruction and activities diverged to create two treatments. In the Stakeholder-Feedback treatment (hereafter shortened to Feedback) students were taught to seek feedback from stakeholders. In the Parallel-Explore condition (hereafter shortened to Explore) students were taught to explore the space of alternatives before settling on a design. At those points, Feedback students sought feedback from others, while the Explore students created multiple designs and sorted among them. Note that we did not attempt to test whether our specific design-thinking curriculum was better or worse than other curricula.

After finishing all the classroom lessons, students in both treatments completed the Poster-Feedback and Photo-Explore assessments, as well as the simpler Apex-Explore assessment. Because neither of the primary assessments (Poster-Feedback and Photo-Explore) resembled nor included content from any of the design-thinking lessons, they are measures of spontaneous transfer of strategies to new learning contexts. The instruction avoided the use of language and specific examples included in the CBAs. The Apex-Explore assessment, however, did build on content the students had covered during an instructional unit (consumers and producers), though it focused on a type of consumer they had not learned about in the classroom. In this respect, Apex-Explore represents a closer transfer than the other two assessments, as the familiar topic may cue students to use the strategies from instruction.

We predicted that students from the two treatments would exhibit contrasting performances on the CBAs. For the Poster-Feedback assessment, we expected that students who completed the Feedback treatment would choose constructive criticism more frequently than students in the Explore treatment would. In contrast, for the Photo-Explore and the Apex-Explore assessments, we expected that students in the Explore treatment would explore the space of alternatives more often than students in the Feedback treatment would.

We were especially interested in whether the patterns of choices and the effects of instruction differed by students' level of academic achievement. One might predict that the instruction would most benefit the highest achieving children because they have been trained to do well in school. This often manifests in behaviors such as avoiding wrong answers and doggedly seeking the single right answer. Alternatively, one might predict that the lower achieving children are more likely to benefit from design-thinking instruction as it provides methods for avoiding common behaviors that may be inadvertently encouraged by “remedial” instruction. Most efforts to change students’ mindsets (e.g., through instruction in growth mindset, role models, or values affirmation) predominantly benefit the students at risk for lower achievement (see Schwartz, Cheng, Salehi, & Wieman, 2016 for a review); this may also be the case for design-thinking instruction.

Methods

Participants. The study recruited the full 6th-grade cohort of a public middle school in California. Of the 237 students who participated, we received consent and useable data for a final sample of *n* = 197. The school is in a high SES neighborhood with 3% of students receiving free or reduced lunch. Of the 170 students who reported their racial-ethnic identities, 71% identified as white, 10% as Hispanic or Latino, 10% as multi-racial, 6% as Asian, and 3% as other. Students participated in the design-based instruction in their math classes (2 weeks), their science classes (1 week), and their social studies classes (3 days). Half of the students did not complete the social studies unit because several teachers needed to cover other content. Analyses indicate that the results did not differ whether students completed the social studies lessons or not, and we do not include completion of the social studies unit as a factor in the following analyses. The school had 12 total math classes. We assigned math classes to either the Feedback or the Explore treatments using stratified random assignment; the stratification variable was the level of mathematics class: regular (3 classes), advanced (6 classes), and double advanced (3 classes). Due to an experimenter error, one class was incorrectly assigned to the Feedback treatment, creating an imbalance in total classes in each treatment. (Feedback treatment, *n* = 101: 2 regular math, 3 advanced math, 2 double advanced math; Explore treatment, *n* = 96: 1 regular math, 3 advanced math, 1 double advanced math.) Students stayed in the same cohort for both math and science classes, therefore maintaining treatment assignment. Students did not stay in exactly the same class grouping for social studies, but the mixing of students was structured such that students remained in the same treatment.

We collected the students' prior year scores on standardized tests in English-language arts (ELA), science, and math. Not all students had scores we could access. ELA, science, and math scores were attained for *n* = 153, *n* = 143, and *n* = 89 students, respectively.

Design. The research had two primary components: (a) Examine the transfer of two design-thinking strategies beyond several weeks of instruction; (b) Determine whether the various assessments index distinct choices that are important for problem solving and learning. The experiment was a two treatment (Feedback vs. Explore) by academic achievement (combined ELA, science, math) by three CBA (Poster-Feedback vs. Photo-Explore vs. Apex-Explore) design. Dependent measures included student choices, as well as measures of performance and learning associated with the assessments.

A researcher and classroom instructor co-taught each class, enabling natural variability across teachers while maintaining treatment fidelity imposed by the researchers (e.g., there could be no mention of any terms or topics used in the CBAs). Each member of the research team and each instructor taught classes in both treatments an equal number of times.

Procedures. The study spanned five weeks, covering three content areas, and involved multiple researchers and teachers leading the activities. Figure 4 captures the high-level design of the experiment.

Our instructional task was to develop a design-thinking curriculum that covered math, social studies, and science, while also teaching design-thinking methods. We created 25 separate lessons across the three content areas. Throughout, we organized the content lessons around a design project and the simple design cycle shown in Figure 5. In math, the projects were determined by the teachers, and included designing a scale model of a house (double-advanced math) or designing a three dimensional box for a new brand of candy (regular and advanced math). In social studies, we created the project to design a process for making fair decisions in school and class. In science, the teachers asked us to cover ecosystems and we created the project to design a game that would teach people about energy transfer through food chains.

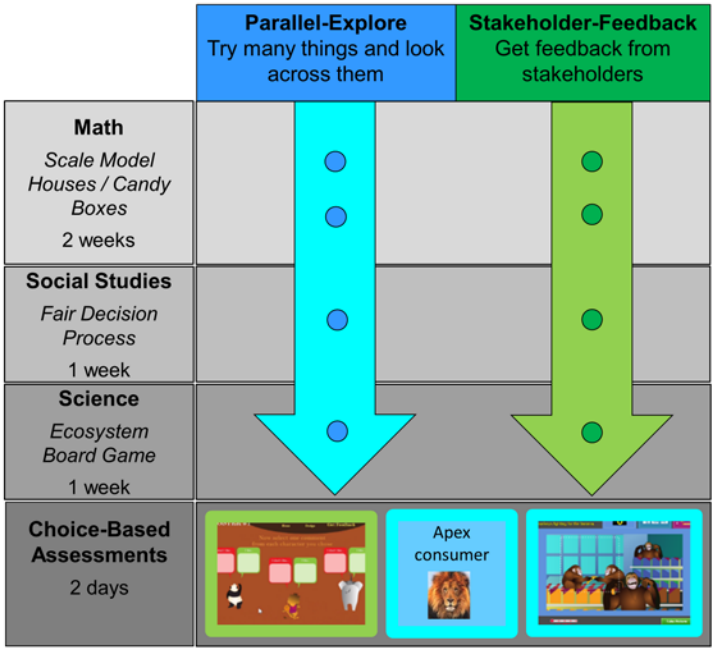


Figure 4. Time course of the experiment. The design projects delivered three subject matter units across both treatments. The dots indicate where the curricular units introduced treatment-specific differences. Both treatment conditions completed all three choice-based assessments.

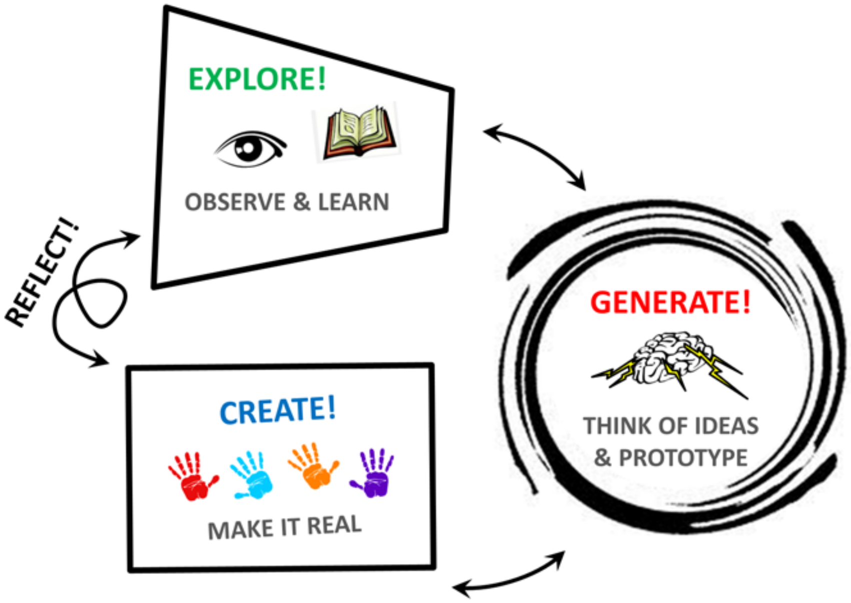


Figure 5. A simplified design cycle used to organize instructional activities into three phases: *Explore! Generate! Create!*

Instructors repeatedly used the visual template of the design cycle in all three content areas, most prominently at the beginning of a day’s lesson. It helped students locate where they were in the design process and how each phase contributed to the larger design process. Through the full instructional period, all students completed a set of common content-focused activities and created their design projects. Critical treatment differences occurred at four key time points, as indicated in Figure 4. There were two during the math unit, and one time each during the social studies and science units. At those points, students in the Feedback treatment had to share their designs with other students and gather their feedback. Students in the Explore treatment instead produced several instances, then compared and discussed them to determine the properties of interest. The instruction for each treatment was explicit about the benefits of its respective design-thinking strategy. For example, in the Feedback treatment, instructors presented several historical disasters that resulted from not seeking feedback, and then facilitated a brief classroom discussion of why people may avoid constructive criticism. In the Explore treatment, instructors led an interactive discussion on the form and function of scissors, where students had to sort multiple pairs of scissors on key features (e.g., blade shape and fulcrum position) and discuss how these design differences could make a pair of scissors better or less suited to its purpose. In general, the instruction tried to provide experience with the benefits of design thinking and included direct explanations of the strategies and their rationale. (Though the specifics of the curriculum are not at test in this study, Appendix A holds descriptions of the instructional sequences to provide a more concrete sense of the design-thinking instruction and the differences between the treatments.) After completing the instruction, all students took the three choice-based assessments on the following two days: (Friday) Poster-Feedback; (Monday) Apex-Explore and Photo-Explore.

Statistical techniques employed in data analyses. In this section, we provide a brief overview of the analytic methods. In the Results, we present the outcomes of the best-fitting models from the analyses. Technical information and complete parameter estimates from these analyses are in Appendix B.

Complete ELA, math, and science achievement data was not accessible for all students. Additionally, several students did not complete all the post-tests for incidental reasons.

Through multiple imputation, as suggested by the editor, we retained a sample size of *n* = 197.[1] Ten imputed datasets (each with *n* = 197) were drawn to estimate achievement and treatment effects and minimize variability from random sampling. For each imputed dataset, a standardized achievement metric (hereafter SAM) was computed as a composite measure of achievement, where SAM is the average of the standardized ELA, math, and science test scores. We used a composite achievement score to reduce the number of fitted parameters, and we did not have *a priori* hypotheses regarding the interactions of specific types of academic achievement and students' choices (see Cutumisu, et al., 2015). All analyses (and corresponding results tables) presented in this paper use the imputed data unless otherwise specified.

The paper includes three main types of statistical analyses. (1) We conducted a multivariate mixed-model regression using both the original data and the imputed data. We designed the model of interest to test the fixed effects of treatment and achievement on the choices in the three CBAs when nested within the random effects of class and student. This random-effects nesting takes advantage of the information that the same student completes each CBA and controls for the effects of intact classes (e.g., some classes or teachers may be generally better performing than other classes). (2) We calculated Pearson correlation coefficients to examine the relations (a) between strategy choices and prior achievement for each CBA, split out by treatment, (b) among strategy choices across three CBAs, and (c) between strategy choices, problem-solving performance, and post-test performance within each CBA. (3) We employed mediation analyses to explore causal relations between strategy choices, problem-solving performance, and post-test learning measures, while controlling for achievement and treatment. More specifically, these examined whether strategy choices flowed through problem-solving performance to affect post-test learning. (At the time of writing, the PROCESS macro for SPSS did not have the capability to include random effects of

[1] Multiple imputation of a given variable uses other available data to predict the value of a missing data point (e.g., using a multiple regression). Given the estimate of the value, the imputation then uses the variance of the existing data to construct a distribution around the value and then randomly samples from that distribution. Each imputation gives a slightly different value for each missing data point due to the random sampling. Further details about the imputation model used for this analysis are available in Appendix B.

class or student within the model. Additional technical details about the mixed-model and mediation analyses appear in Appendix B.)

Results

The results come in two main sections. The first section focuses on the design-thinking strategy choices and their generalization to other contexts and measures of achievement. It determines whether instruction affected student choices in a transfer context, as well as the relations among strategies and academic achievement. The second main section focuses on the problem-solving performance and learning measures. It explores how strategy choices within the CBAs affect performance and learning from the CBAs.

The generalization of distinct design-thinking strategies. A major goal of the study was to answer the question of whether the design-thinking strategies acquired through instruction transfer beyond that instruction. The three CBAs (Poster-Feedback, Photo-Explore, and Apex-Explore) served as the transfer context. These assessment environments are a reasonable test of spontaneous transfer. They do not look like what the students completed during their instruction. They use new topics. They use tablet computers, which were not part of the instruction. The test of transfer was moderately far, but not so far as to be tasks that do not involve design or occur beyond the boundaries of classrooms. The challenges posed in the two main CBAs both required design (Poster-Feedback and Photo-Explore). Moreover, the assessments occurred in class and were administered by the research team.

The critical test is whether the different instructional interventions have selective treatment effects on choice behaviors. Students in the Feedback treatment should be more likely to choose more constructive criticism (“I don’t like”) during the Poster-Feedback CBA than their counterparts in the Explore treatment. The Feedback students were taught to seek stakeholder feedback, whereas the Explore students were not; hence the Explore students serve as controls in this case. In contrast, the Explore students should be more likely to explore multiple camera settings in Photo-Explore and examine more animals in Apex-Explore than students in the Feedback treatment. The Explore treatment students were taught to explore multiple instances before settling on a final design, whereas their Feedback counterparts were not; hence the Feedback students serve as controls in this case.

A second set of questions is how prior achievement interacts with the use of the design-thinking strategies. In particular, would lower-achieving students benefit more from instruction, as has been found in other interventions that target psychological barriers to success in school (for a brief review, see Schwartz, Cheng, Salehi, & Wieman, 2016) and would strategy use correlate with measures of prior learning (i.e., standardized achievement)? A final, related question is whether the CBAs measure distinct strategies or whether they are a subset of a more generalized construct responsible for individual differences in achievement, such as mindset.

The following mixed-model analysis addresses both the first and second set of questions simultaneously. To see the implications of this analysis more plainly, we follow the main analysis with two subsections that present simple correlations and additional relevant data to help make the points more directly.

Strategy choices can transfer. Figure 6 shows the choices for students who have been taught design-thinking strategies relevant to a given CBA, as compared to those students who did not receive instruction relevant to the CBA (the "control" group). The plots indicate the average choice scores binned by the low, middle, and top thirds of SAM.[2] In the respective control conditions (represented by dotted lines), lower-achieving students made relatively fewer choices to seek constructive criticism of their posters, made fewer explorations of camera settings, and checked fewer animals than their higher-achieving peers. However, a few hours of instruction in relevant design-thinking strategies reduced the difference between lower- and higher-achieving students in the treatment group for each assessment (represented by solid lines).

[2] Note the bin width values were determined by the distribution of SAM from the imputations, not the original data set.

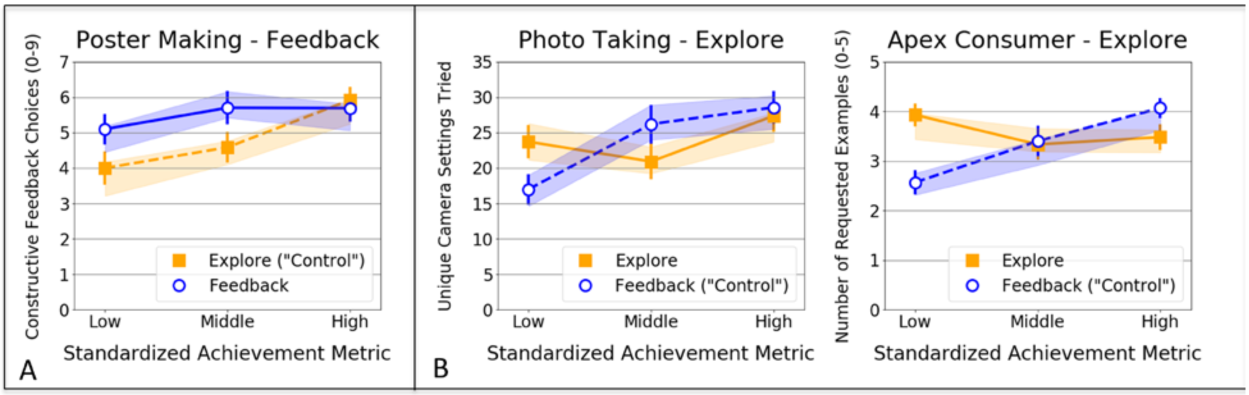


Figure 6. CBA choices as a function of standardized achievement metric (SAM) and treatment for A) Poster Making-Feedback and B) Photo Taking-Explore and Apex Consumer-Explore. Solid (treatment) and dashed (control) lines indicate observations from the original dataset. Shaded bands represent the spread of the ten imputed datasets. The width of the bands represent a standard error in each direction.

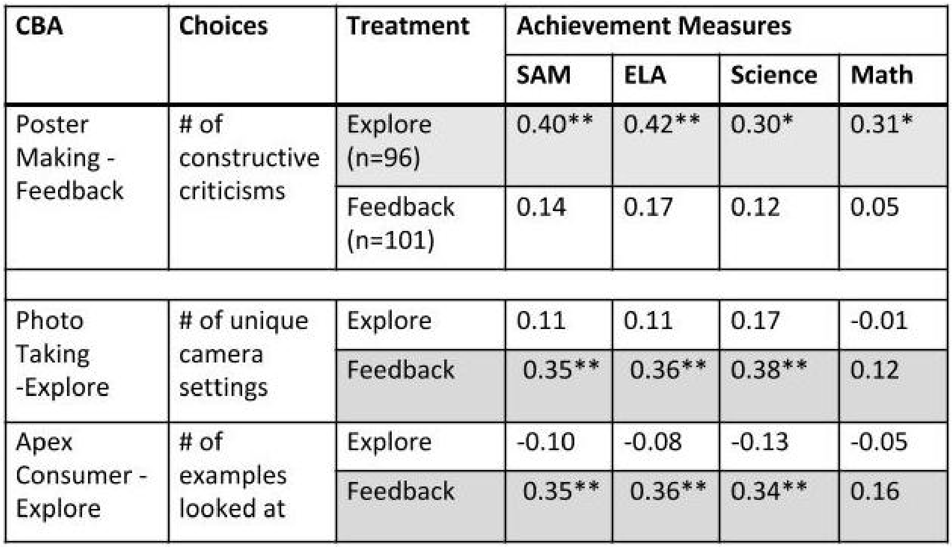
The mixed-model analysis confirms the patterns in Figure 6. First, we describe the results using the original, non-imputed data (i.e., students for whom we had complete in-game choice data), then we present the statistics with the imputed data. Using the non-imputed data, there was a main effect of prior achievement on CBA choices, F(3,197.5) = 5.37, *p* = 0.001. This effect can be seen in the generally upward slope of the lines from lower- to higher-achieving students in all three sub-panels of Figure 6. There was also a significant interaction between achievement and treatment, F(3, 182.5)=6.51, *p* < 0.001. The lower achieving students drive the interaction. Figure 6 exhibits a gap between the lower-achieving control and treatment students for all three CBAs. There was no stand-alone effect of treatment, F(3, 29.4) = 0.88, *p* = 0.46.

The results were similar using the imputed data. There was an effect of achievement across all ten imputations; p-values ranged from *p* < .001 to *p* < 0.009. There was no effect of treatment; p-values ranged from *p* = 0.295 to *p* = 0.376. There was an interaction between achievement and treatment; p-values ranged from *p* < 0.001 to *p* < 0.004.

The gap between the lower-achieving control and treatment students flips between the Feedback and Explore CBAs. For Poster-Feedback, the low-achieving Feedback students do better than their Explore counterparts do. For Photo-Explore, the low-achieving Feedback students to worse than their Explore counterparts do. To test this crossover effect, we use the pooled imputation results to compare parameter estimates of the treatment by achievement interaction for each CBA. (By subtracting two parameter estimates and dividing by their pooled standard errors, we obtain a t-value of the difference.) Setting α = 0.05, the comparisons indicate that the effect of treatment by achievement on Poster-Feedback is significantly different from both Photo-Explore (*t* = 2.08) and Apex-Explore (*t* = 3.17). The effect of treatment by achievement does not significantly differ between Photo-Explore and Apex-Explore (*t* = 1.02). The fact that the interaction effects on Photo-Explore and Apex-Explore are not significantly different suggests they are measuring the same construct (i.e., the choice to explore).

Strategies and prior achievement are related. The preceding analysis indicates that choices within a CBA are associated with prior achievement and that strategy instruction may have "broken" this relationship for a portion of students who received strategy instruction relevant to the CBA. To elaborate on this finding, Table 2 presents the simple correlations between each measure of prior achievement and the strategy choice measures from each CBA, separated by treatment. The shaded portions indicate correlations for the respective control students (i.e., students who did not receive instruction relevant to the choices measured by a particular CBA.)

Table 2. *Correlations between choices and achievement measures broken outby CBA and treatment. Shading indicates respective control for each assessment. ** p < 0.01; * p < 0.05.*



CBA	Choices	Treatment	Achievement Measures			
			SAM	ELA	Science	Math
Poster Making - Feedback	# of constructive criticisms	Explore (n=96)	0.40**	0.42**	0.30*	0.31*
		Feedback (n=101)	0.14	0.17	0.12	0.05
Photo Taking -Explore	# of unique camera settings	Explore	0.11	0.11	0.17	-0.01
		Feedback	0.35**	0.36**	0.38**	0.12
Apex Consumer - Explore	# of examples looked at	Explore	-0.10	-0.08	-0.13	-0.05
		Feedback	0.35**	0.36**	0.34**	0.16

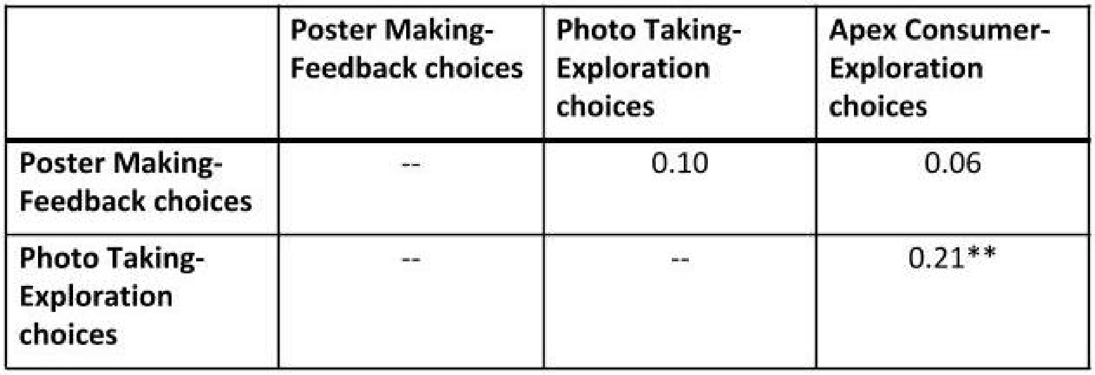
The Poster-Feedback assessment (top rows) was designed to measure the choice to seek critical feedback. The Explore treatment students, who did not receive instruction on seeking feedback, exhibit a significant correlation between Feedback choices and all standardized measures of achievement. Presumably, the CBA captures historical patterns of behavior that influenced the students' prior learning and hence their achievement scores. In contrast, for the Feedback students, we intervened to change those behaviors, so the choices they made in the CBA no longer reflect prior patterns of behavior that influenced their academic achievement.

For the Photo-Explore and Apex-Explore CBAs, the same pattern occurs in reverse. The Feedback treatment students, who did not receive instruction on exploring a space of alternatives, exhibit significant correlations between their prior achievement scores and exploration choices. Explore treatment students, who did receive exploration instructions, do not exhibit correlations between achievement and their choices in the games.

The correlations between choices and achievement among the untreated students are statistically significant, but practically moderate. Nevertheless, it is striking that choices during a 10-minute online interaction can account for 10% to 15% of the variance in standardized test performance, especially when those choices are not directly related to the content of those tests. Moreover, the correlations are largely present across achievement tests that measure distinct domains of knowledge. The one exception is the lack of correlation between math achievement and the two Explore CBAs. We do not have a strong explanation for why math achievement correlates with Poster-Feedback but not Photo-Explore or Apex-Explore. Perhaps these students' math instruction had been proceduralized such that there was minimal opportunity for student exploration behaviors to affect their past learning.

The CBAs measure separate constructs. Both the mixed-model and correlational analyses in the previous sections suggest that instruction could change the choice to use one design-thinking strategy without changing the other. This selective treatment effect on choices implies that the choices within a CBA are not the result of a more general construct such as IQ or mindset. To more directly examine this point, Table 3 shows the cross-correlations between Poster-Feedback, Photo-Explore, and Apex-Explore.

Table 3. *Correlations between choice behaviors across the three CBAs (n = 197).* ** p < 0.01



	Poster Making-Feedback choices	Photo Taking-Exploration choices	Apex Consumer-Exploration choices
Poster Making-Feedback choices	--	0.10	0.06
Photo Taking-Exploration choices	--	--	0.21**

The frequency with which students chose constructive criticism in Poster-Feedback is uncorrelated with the frequency of exploring in either Photo-Explore or Apex-Explore. In contrast, there is a significant correlation between Photo-Explore and Apex-Explore. These results indicate that Photo-Explore and Apex-Explore may be measuring a similar construct that is different from what Photo-Feedback is measuring.

One concern with this interpretation is that we intervened with instruction targeting one type of CBA but not the other for each student. It is possible that had we not intervened, choices in the Poster-Feedback and Explore CBAs would have exhibited stronger correlations. One way to address this concern is to take advantage of the fact that the higher-achieving students did not exhibit a strong response to the instruction. This group of students may determine whether a correlation exists between the choices for relatively untreated students. We used a median split on the SAM achievement measure and re-computed the correlations for the lower- and higher-achieving students separately. Taken separately, the lower-achieving children exhibit a low correlation between Poster-Feedback and Photo-Explore (*r* = 0.08). The higher-achieving students also do not exhibit a meaningful correlation (*r* = 0.05). These two CBAs appear to be measuring distinct choices rather than a single underlying ability or mindset.

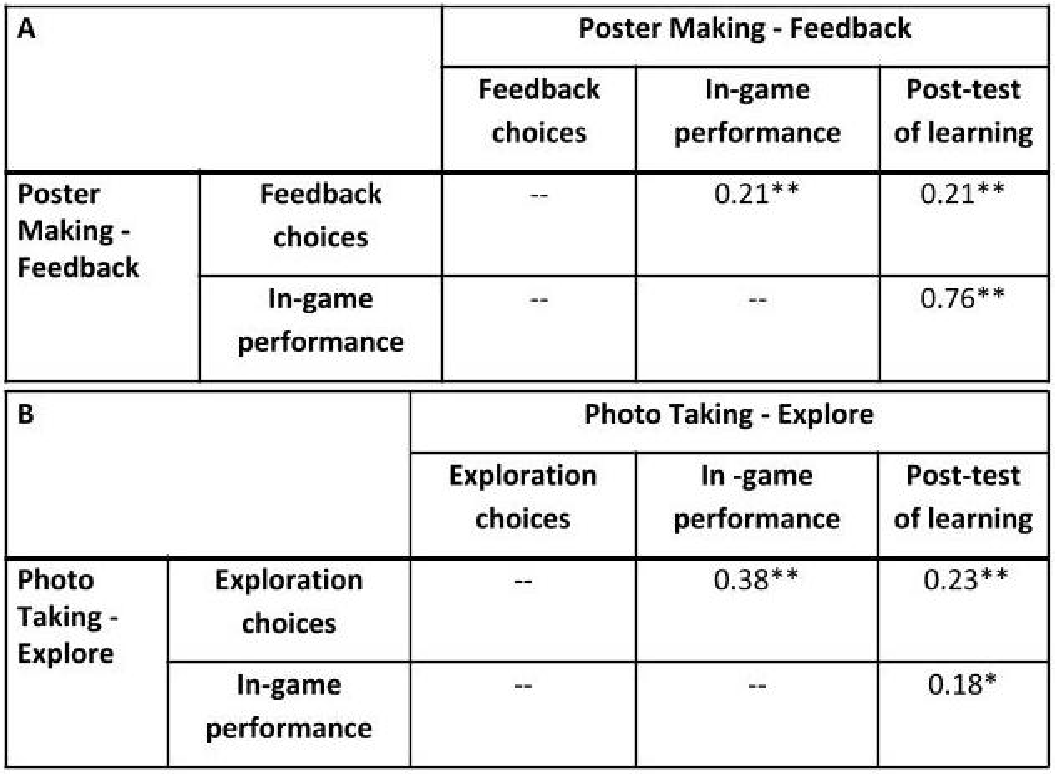
The effect of choices on in-game performance and post-test measures of learning. This section considers whether the choices to seek constructive criticism and explore a space of alternatives are good choices for problem-solving performance and learning. CBAs are designed to include information that students can learn during the assessment. In Poster-Feedback, there is information about graphic design principles. In Photo-Explore, there is information about the effects of camera settings (focus, shutter speed, and aperture). For each of these two CBAs, there was a measure of problem-solving performance within the environment, as well as a measure of learning, in the form of a brief post-test outside the environment.[3]

Poster-Feedback includes an intelligent system that evaluates the formal graphic features of each poster (e.g., legible font). To measure in-game problem-solving performance, we compared the quality of the first and last poster of each student. This created a graphic design improvement score, which we call Poster Quality Gain. Photo-Explore includes an engine that evaluates the exposure and blurriness of each photo. To measure in-game problem-solving performance in this CBA, we created a total photo score by summing across a student's four submitted photos. We call this Aggregate Photo Quality. (We cannot use a gain score for the Photo-Explore CBA because students did not have access to all the camera settings in the early levels.) To measure out-of-game learning, we use the results from the brief post-tests appended to each assessment environment.

[3] We did not develop performance or learning measures for Apex-Explore.

Choices correlate with performance and learning. Panel A of Table 4 shows that seeking constructive criticism in Poster-Feedback correlates with better in-game problem-solving performance and post-test learning. Panel B shows that exploring a space of alternatives in Photo-Explore correlates with better in-game performance and post-test learning. The associations are small to moderate, which is to be expected. The choices are not direct measures of what students learned, but rather, what information they chose to learn from.

Table 4. *Correlations between choices and learning measures within (A) Poster-Feedback and (B) Photo-Explore (n= 197).* ** p < 0.01; * p < 0.05.



A		Poster Making - Feedback		
		Feedback choices	In-game performance	Post-test of learning
Poster Making - Feedback	Feedback choices	--	0.21**	0.21**
	In-game performance	--	--	0.76**

B		Photo Taking - Explore		
		Exploration choices	In -game performance	Post-test of learning
Photo Taking - Explore	Exploration choices	--	0.38**	0.23**
	In-game performance	--	--	0.18*

Choices mediate performance and learning. In the simple correlations of Table 4, there is the possibility that a latent variable, such as general school achievement, is causing both choices and outcomes. To address this possibility, we employ a mediation analysis to control for the effects of achievement and treatment to see if there is still an association between choices, in-game problem-solving performance, and post-test learning. (See Appendix B for technical details of the analysis.)

Figure 7 shows the results of the mediation analysis for Poster-Feedback. Key information is displayed in the links that connect the various measures.

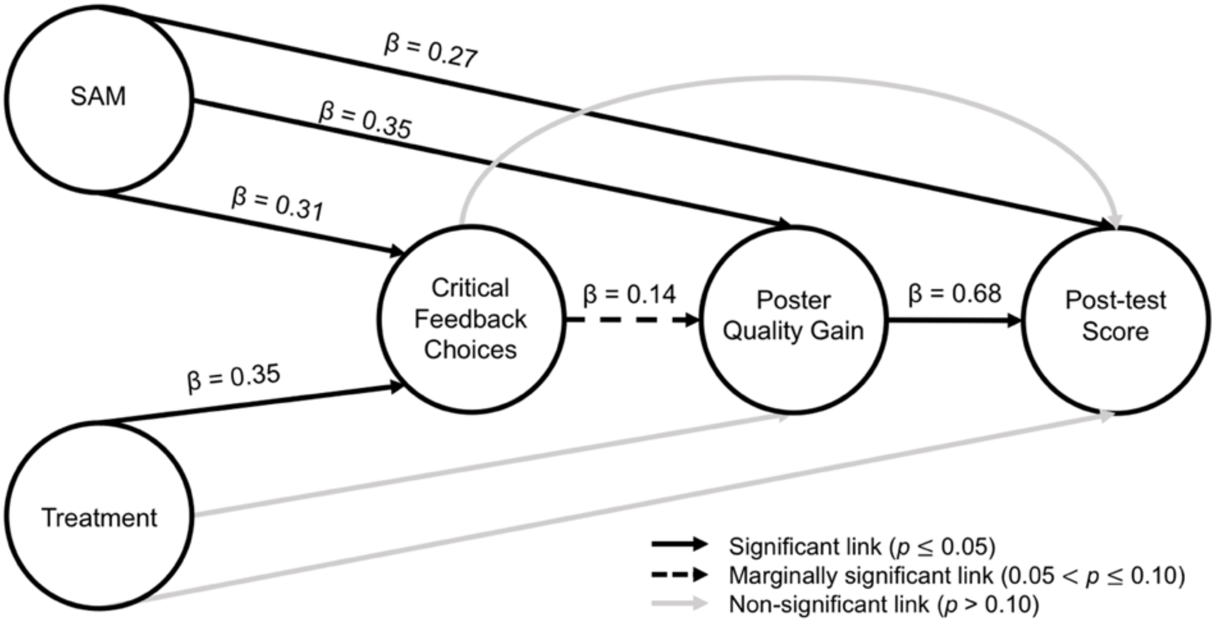

SAM
Treatment
Critical Feedback Choices
Poster Quality Gain
Post-test Score
β = 0.27
β = 0.35
β = 0.31
β = 0.35
β = 0.14
β = 0.68
Significant link ($p \leq 0.05$)
Marginally significant link ($0.05 < p \leq 0.10$)
Non-significant link ($p > 0.10$)

Figure 7. Mediation analysis of Poster-Feedback performance and learning outcomes, including the standardized achievement metric (SAM) and treatment condition as independent variables.

The dashed line between Critical Feedback Choices and Poster Quality Gain indicates a marginally significant association between strategy choices and problem-solving performance improvement. The β = 0.14 means that for every standard deviation increase in choosing critical feedback, there is a 0.14 standard deviation increase in the Poster Quality Gain score (it may be interpreted as $r = 0.14$). In turn, Poster Quality Gain is strongly associated with learning of graphic design principles, as measured by the post-test. In this case, the effect of choices runs through performance gains within the game, which drives better results on the post-test. Overall, the evidence that choosing critical feedback improves learning is modest when taking into consideration prior achievement. There are multiple possible interpretations. One interpretation is that the choice to seek constructive criticism mildly affects gains in learning in this context. A second interpretation is that the measures of in-game improvement and learning are poorly designed. A third interpretation is that choosing critical feedback is one of the reasons students' exhibit better achievement in the first place, and that achievement is actually explaining some of the variance that feedback choices would otherwise explain. One possible solution to evaluating these alternatives is to conduct a study where students are assigned to different frequencies of constructive criticism to determine if more constructive criticism causes more learning. To succeed in this approach, it will be necessary to determine whether choosing versus receiving constructive criticism have differential impacts on learning (for an approach to this problem, see Cutumisu & Schwartz, 2018).

Figure 8 shows the results for Photo-Explore. In this case, the choice to explore more unique camera settings significantly predicts the problem-solving performance outcome (photo quality). However, the quality of the photos does not significantly predict the post-test

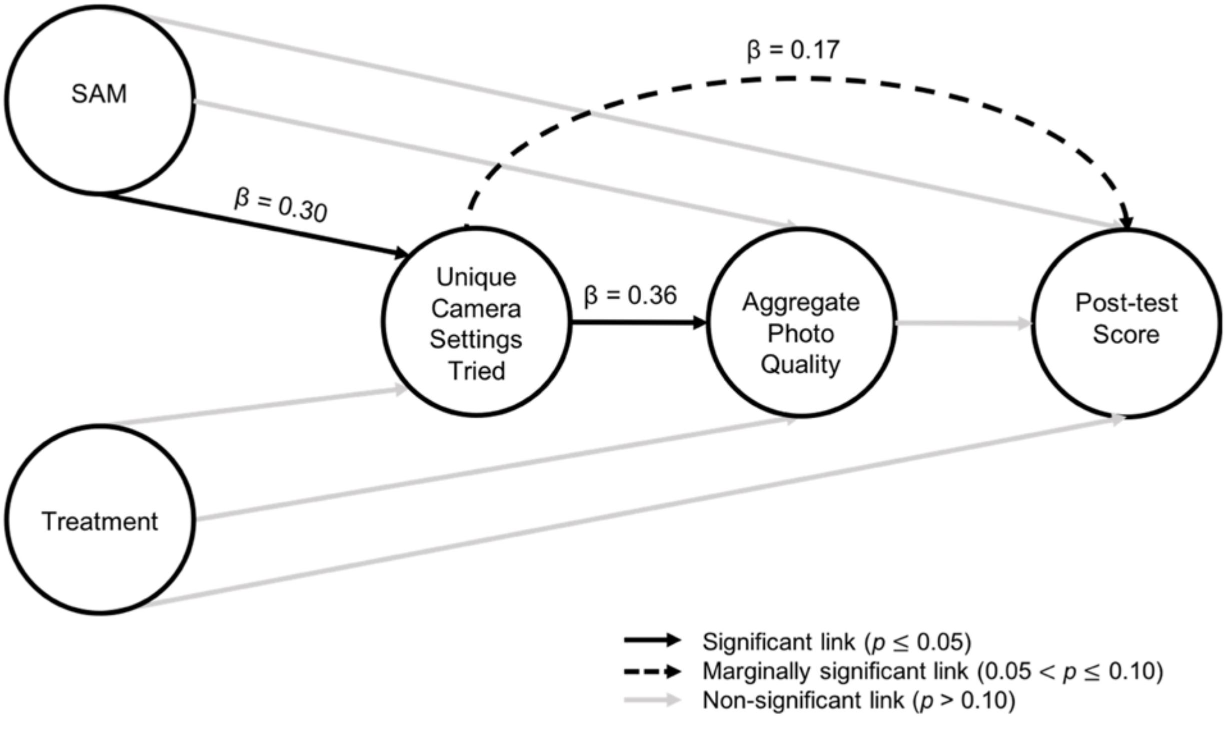

SAM
β = 0.17
β = 0.30
Unique Camera Settings Tried
β = 0.36
Aggregate Photo Quality
Post-test Score
Treatment
Significant link (p ≤ 0.05)
Marginally significant link (0.05 < p ≤ 0.10)
Non-significant link (p > 0.10)

Figure 8. Mediation analysis of Photo-Explore performance and learning outcomes, including the standardized achievement metric (SAM) and treatment condition as independent variables.

measure of learning. This makes some sense in that the average quality of photos does not capture learning from the start to the end of the game, and therefore, it is a sub-optimal measure to correlate with the post-test. The choice to explore more settings does show a marginally significant direct effect on post-test scores, indicating that CBA choices have a modest impact on learning as measured outside of the game. Thus, as was the case with Poster-Feedback, the evidence that the exploration choices improve learning is modest, though promising. We did not vet the measures of in-game problem-solving performance and post-test learning prior to using the assessments in this study, so it is conceivable that strengthening the measures of performance and learning would improve their association with the choice to explore a space of alternatives.

General Discussion

Empirical Summary

The research discussed in this article offers a relatively strong test of the spontaneous transfer of two broadly applicable strategies: seeking constructive criticism and exploring a space of alternatives. After design-thinking instruction in one or the other of the two strategies, 6th-grade students completed choice-based assessments (CBAs). The assessments offered activities without an obvious connection to the context, topic, or look of the original instruction. This lack of similarity to the conditions of learning is what defines the CBAs as a measure of transfer. They are unusual transfer instruments because they attempt to measure students' preparation for future learning (Bransford & Schwartz, 1999). They include resources for learning new information relevant to solving the activity challenge, but there are no explicit task demands to learn the content or to use the learning strategies they may have previously learned. Thus, they measure whether students spontaneously choose to use the strategies for learning as they might do in the future.

Using the CBAs, we found that the two design-thinking strategies are behaviorally measurable, can be taught, are correlated with prior achievement, and may modestly improve problem solving and independent learning of new content. Lower-achieving students exhibited the greatest evidence of benefiting from and transferring the design-thinking instruction, as the higher-achieving students tended to employ the two strategies regardless of their instructional intervention.

In more detail, the CBAs yielded promising, but not definitive, evidence on whether the choice to use the two design strategies improves learning. The Poster-Feedback CBA measured students' choices to seek constructive negative feedback; we found that this choice was marginally associated with better problem solving during the assessment, as measured by greater poster improvement. In turn, poster improvement was associated with increased learning of graphic design principles, as measured by higher performance on a separate post-test. The Photo-Explore assessment measured whether students explored a space of possibilities before settling on a solution. This was also associated with improved problem solving, as measured by photo quality within the CBA, as well as marginally associated with increased learning, as measured by a post-test.

Students also completed Apex-Explore. Choices to see more examples in Apex-Explore correlated with the exploring more camera settings Photo-Explore, but did not correlate with seeking more constructive criticism in Poster-Feedback. This lends some convergent and divergent validity to the claim that Photo-Explore measured students' choices to explore a space.

Choices on the two primary CBAs – Photo-Explore and Poster-Feedback – correlated with prior standardized achievement in aggregate (SAM) as well as each achievement score in isolation. This raises the possibility that they are measuring the same underlying construct, such as IQ. However, the interaction effect from the mixed-model analysis, as well as the correlational analyses presented in Tables 2 and 3, all indicate the CBAs are measuring different psychological constructs. We propose that these choices reflect strategies that people may or may not adopt to help overcome natural tendencies that can interfere with problem solving and learning. While there is a theoretician's urge to offer a single cognitive or dispositional construct that explains why people do or do not make good choices, these results suggest it may be more profitable to think in terms of heuristic strategies and methods rather than inherent dispositions or traits.

The results provide an important demonstration of the ability to measure two key aspects of design-based learning strategies, and learning strategies in general. First, they measure whether students choose to apply them in new situations. Second, they ideally measure whether the strategies improve problem solving and learning.

Towards Design Principles for Choice-Based Assessments

If we accept that an important goal of education is to prepare students for a changing future, then we would ideally equip them with strategies for learning on their own. This entails an emphasis on the processes of learning, and not just mastery of the knowledge delivered in school. The need to focus on process appears in reports by the National Science Foundation (Friedman, 2008) and the National Research Council (2009). It is also formalized in curricular frameworks for education including the Common Core and Next Generation Science Standards, both of which incorporate disciplinary practices in their list of competencies (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; NGSS Lead States, 2013). These curricular goals require new kinds of assessments that can directly measure processes of learning relevant to specific domains and activity contexts. Such goals also require a reconceptualization of the psychology of learning competencies.

As we argue throughout this article, it is not enough for people to know these competencies; they also need to choose to use them. People know that exercise is a good method for increasing healthfulness, but they may still choose to be sedentary. One way to re-conceptualize learning competencies is to recognize that they always occur in a context of competing alternatives rather than simply being right or wrong. People may choose to spend their free time reading or watching television, instead of exercising. In the context of design, we had students choose between the competing alternatives of praise versus constructive criticism and settling on an early solution versus exploring a space of alternatives.

Our general approach to the creation of choice-based assessments is to consider why people might not engage in a problem-solving or learning process. In the context of design, we aimed at the natural tendency towards early closure, where people can unwittingly settle for a sub-optimal solution. Seeking praise is a way to speed early closure because it signals one has performed satisfactorily. In contrast, choosing constructive criticism is a way to learn what needs improvement. Seeking constructive feedback and exploring the space are two strategies that help prevent the natural tendency to early closure.

If we were trying to teach people additional learning skills, such as self-explanation while reading (Chi et al., 1994), we would begin from the assumption that there are reasons people choose not to self-explain, even if they know how to do it. For example, self-explanation is cognitively demanding and relatively slow. To assess whether students self-explain, we would create an environment where they could “level-up” without self-explaining, even though self-explaining could provide some benefits for performance. With respect to instruction, we would help students understand the reasons they might choose not to self-explain, and ideally, provide them with convincing evidence that it is worth the effort and time. Thinking about choice in the design of instruction and assessment leads to a focus on the alternatives. We offer a hypothesis for future research: Instruction of a particular learning strategy will be more effective when it explicitly helps students address why they might naturally choose against using that strategy (cf. Oppezzo & Schwartz, 2013). For example, for scientific inquiry, we would warn students of the tendencies to over rely on their beliefs or intuitions at the expense of reasoning from evidence.

The current CBAs were designed with choice alternatives that reveal natural tendencies that may lead people to choose against the desired learning strategy. More generally, we offer three principles for the design of choice-based assessments.

Preparation for future learning principle. We assume that one major argument for process and dispositional instruction is that it will help students learn in the future. Therefore, assessments should include opportunities to learn so we can determine if students have been prepared to learn by making productive choices. CBAs are an instance of dynamic assessments that evaluate what students can learn as part of the testing process (Feuerstein, 1979).

Free choice principle. We want to model the relatively choice-rich environment in which students will learn after they leave school. These environments are less scripted than school, and people need to decide what, whether, and how to learn. To get a fair measure of people’s unguided decisions and behaviors, assessments cannot fully drive student choices

(Schwartz & Arena, 2013). For instance, in our computerized assessments, students can progress to new challenges, no matter what learning choices they make.

Typical performance principle. In an assessment context, students often display “maximal” effort and test-taking behaviors (Sackett, Zedeck, & Fogli, 1988; Klehe & Anderson, 2007), which may not reflect students’ typical performance. To get a better measure of the choices that students would normally take, we create short and friendly assessment environments that do not feel like a test and do not drive students towards compliance. (One advantage of using friendly designs is that CBAs can be deployed in settings of informal learning.) Assessments of typical behaviors have also been called “stealth assessments” (Shute & Ventura, 2013) because students are unaware they are being tested, let alone what is being tested. There are ethical considerations when testing people who do not know they are being evaluated (Schwartz & Arena, 2013). These considerations are especially important for high-stakes evaluations, which the current assessments are not. Our primary practical goal in creating these assessments is to provide formative feedback to instructional programs about whether they are meeting their intended goals of helping students develop productive learning strategies.

Limitations and Future Research

One sign of a useful measurement is that it generates new research questions that go beyond its initial use. This appears to be the case here. One striking result was that the design-thinking curriculum had the strongest effects on the lower-achieving students. When we began the study, several teachers and colleagues thought that the highest-achieving students needed the design-thinking lessons because these students always want to know they are doing well in school and avoid criticism. This was not the case. The highest-achieving students chose the most constructive criticism regardless of instruction. One possible explanation is that the design activities were atypical of school, and therefore these students adaptively shed their typical classroom behaviors. However, this explanation is inconsistent with the finding that the choices correlated with historical achievement in school. The correlations suggest that students were previously using these strategies (e.g., seeking constructive criticism) in the context of typical school assignments. Perhaps the students applied these choices outside of the classroom, for example, during homework. Another explanation for the mistaken prediction by our colleagues is that these kinds of choices may be harder to notice in normal school interactions. A good measurement can reveal what may not be easily observable in the flow of daily classroom events.

Four questions stand out for future research. The first asks why the lower-achieving students in the respective “control” conditions scored lower on the strategy choice measures than the higher-achieving students. One proposal is that lower-achieving students typically receive a steady diet of implicit criticism about their performances, and therefore, they shy away from asking for even more criticism in Poster-Feedback. This would not explain why the lower-achieving students were also less likely to explore the space of alternatives in Photo-Explore, which seems less relevant to ego threat. A second proposal is that the lower-achieving students had always received strong guidance on exactly what they were supposed to do, which is often the case for lower-achieving students. In turn, this reduced their opportunities to make choices about their problem solving and learning over the years. A third proposal is that the lower-achieving students had fewer opportunities to engage in structured informal learning experiences, such as science camp, that would foster strategies for learning outside of typical classroom instruction. A fourth proposal is that lower-achieving students, in general, simply engage less in learning-relevant processes. However, this “desultory student” explanation is unsupported by the data, which showed no correlation between the behaviors in the two CBAs. The lack of correlation makes it unlikely there is an umbrella disposition or mindset that hinders the lower-achieving students in general, at least in the context of these design-thinking strategies.

The second question asks why only the lower-achieving students exhibited an effect of the instruction. This finding has precedent across many of the social-psychological interventions that attempt to improve learner’s self-attributions. These studies often find that the interventions, whether growth mindset, values affirmation, or social modeling, primarily take hold for the subset of students who are at risk for under-achieving in school. One typical explanation is that the assessments exhibit a ceiling effect, so that higher-achieving students have no room to show improvement. In the current study, this was not the case. There was room for improvement for the students in the middle achievement rank, and they did not appreciably change their choices across all three CBAs because of the instruction. A second explanation for these results is that the lower-achieving students were more receptive to learning how to learn, at least in the novel context of design. For them, it was a fresh start, whereas for the more successful students it was part of their regular curriculum, and so they did not take the lessons to heart (children in advanced classes frequently ask, “Does this count for our grade?”).

A third question asks what types of design-thinking instruction are effective. The research design did not address this question. It is possible that had we provided better lessons, the higher-achieving students would have adopted the strategies more frequently. On the other hand, it is possible that a different method for teaching design thinking would have failed to reach the lower-achieving students. Now that we have the template for creating choice-based assessments, it will be possible to investigate different instructional models and features.

The final question addresses the generalizability of these results to other populations and across other settings. Children in this study were from a high SES community with relatively low racial diversity, and they participated within the confines of a classroom setting. Would the effects generalize to other populations and learning settings such as summer camp? We need more research on questions of reach because of the potential implications for how to help all children prepare for a dynamic future.

Conclusion

The content of assessments merits intense consideration, because assessments evaluate and enforce what is important to learn. Current assessments largely emphasize what knowledge people have accrued, typically in the form of facts, concepts, and algorithms. Given an unknown future, we would offer that it is vital to assess whether people are prepared to make productive choices that increase their chances of learning, so it is possible to teach towards this goal and know if one is doing so effectively. By our analysis, one major, yet under developed, goal of assessment is to determine if students will choose to use specific strategies, and by implication, a major goal of instruction is to influence students' choices (Schwartz & Arena, 2013). Our use of choice differs from most educational research, where student choice is taken as a variable that drives motivation and students' sense of autonomy (Iyengar & Lepper, 1999; Ryan & Deci, 2000). Here, we take learning to choose well as a behavioral outcome of effective instruction, and therefore, choices are part of what we should teach and assess. The current research presents one example of how to evaluate whether students are gaining productive choices that they use in new situations. Thus far, it appears that seeking constructive criticism and exploring a space of alternatives can modestly improve problem solving and learning, and that children can learn to choose these strategies in a novel context where there is nobody telling them what to do.

Acknowledgements

We would like to thank the students, teachers, and administrators, particularly Principal Erik Burmeister, of Hillview Middle School, Menlo Park, California. Without their support and enthusiasm, this study would not have been possible.

References

Acredolo, C. & Horobin, K. (1987). Development of relational reasoning and avoidance of premature closure. *Developmental Psychology*, *23*(1), 13.

Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *Journal of General Psychology, 54*, 279-299.

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

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In Lester, J.C., Vicari, R.M., Paraguaçu, F. (Eds.), *Intelligent Tutoring Systems*, 3220, (pp. 531–540). Springer, Heidelberg, Germany.

Bamberger, Y., & Cahill, C. (2013). Teaching design in middle-school: Instructors' concerns and scaffolding strategies. *Journal of Science Education and Technology, 22*(2), 171-185.

Barron, B. J., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., Bransford, J. D., & CTGV. (1998). Doing with understanding: Lessons from research on problem- and project-based learning. *Journal of the Learning Sciences, 7*, 271-312.

Biswas, G., Jeong, H., Kinnebrew, J.S., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning, 5*(2): 123-152. DOI: 10.1142/S1793206810000839

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322-330. https://doi.org/10.1016/j.cognition.2010.10.001

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education*, *24*, (pp. 61-101). Washington DC: American Educational Research Association.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher, 10*(2), 14-21

Brown, T. & Wyatt, J. (2010). Design thinking for social innovation. *Stanford Social Innovation Review,* 8(1), 30-35.

Carroll, M., Goldman, S., Britos, L., Koh, J., Royalty, A., & Hornstein, M. (2010). Destination, imagination and the fires within: Design thinking in a middle school classroom. *International Journal of Art & Design Education*, *29*(1), 37-53. DOI**:** 10.1111/j.1476-8070.2010.01632.x

Chase, C., Chin, D. B., Oppezzo, M., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology, 18,* 334-352. DOI: 10.1007/s10956-009-9180-4

Chi, M. T., Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439-477.

Chi, M., VanLehn, K., & Litman, D. (2010). Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In: V. Aleven, J. Kay, & J. Mostow (Eds), *Intelligent tutoring systems. ITS 2010. Lecture Notes in Computer Science, 6094,* (pp. 224-234). Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-13388-6_27

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*(5925), 400-403.

Conlin, L.D., Chin, D.B., Blair, K.P., Cutumisu, M., & Schwartz, D.L. (2015, June). Guardian angels of our better nature: Finding evidence of the benefits of design thinking. *Proceedings of the 2015 Meeting of the American Society of Engineering Education*, Seattle, WA. DOI: 10.18260/p.24165

Couzijn, M. (1999). Learning to write by observation of writing and reading processes; effects on learning and transfer. *Learning and Instruction, 9*(2), 109-142. DOI: 10.1016/S0959-4752(98)00040-1

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 revise. *Journal of Learning Analytics*, *2,* 49-71. DOI: 10.18608/jla.2015.21.4

Cutumisu, M., Blair, K. P, Chin, D. B. & Schwartz, D. L. (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, 27(2).* DOI:10.1007/s40593-016-0137-5.

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

Csikszentmihalyi, M. & Getzels, J. (1970). Concern for discovery: An attitudinal component of creative production 1. *Journal of Personality 38*(1), 91-105.

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science*, *333*(6045), 959-964. DOI: 10.1126/science.1204529

Dow, S. P., Glassco, A., Kass, J., Schwarz, M., Schwartz, D. L., & Klemmer, S. R. (2010). Parallel prototyping leads to better design results, more divergent creations, and self-efficacy gains. *ACM Transactions on Computer-Human Interaction*, *17*(4). DOI: 10.1145/1879831.1879836

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 109-143). New York: Psychology Press.

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

Fancsali, S. E. (2014). Causal discovery with models: Behavior, affect, and learning in cognitive tutor algebra. *Proceedings of the 7th International Conference on Educational Data Mining*, 28-35, London, UK.

Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore, MD: University Park Press.

Finkelstein, S. R., & Fishbach, A. (2012). Tell me what I did wrong: experts seek and respond to negative feedback. *Journal of Consumer Research*, *39*(1), 22-38. **DOI:** https://doi.org/10.1086/661934

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

Fuchs, L.S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C.L., Owen, R., & Schroeter, K. (2003). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology, 95*(2): 306-315.

Goldman, S., & Kabayadondo, Z. (Eds.). (2017). *Taking design thinking to school: How the technology of design can transform teachers, learners, and classrooms*. New York: Routledge.

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

Hoskins, B., & Fredriksson, U. (2008). Learning to Learn: What is it and can it be measured. *Joint Research Centre Technical Report JRC*, *46532.*

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.

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

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

Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67–72.

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

Koh, J. H. L., Chai, C. S., Wong, B., & Hong, H. Y. (2015). *Design thinking for education: Conceptions and applications in teaching and learning*. Singapore: Springer. DOI: 10.1007/978-981-287-444-3

Kolodner, J.L., Crismond, D., Fasse, B.B., Gray, J., Holbrook, J., & Puntembakar, S. (2003). Putting a student-centered Learning By Design^{TM} curriculum into practice: Lessons learned. *Journal of the Learning Sciences, 12*(4), 495-547. DOI: 10.1207/S15327809JLS1204_2

Kulik, J.A., & Kulik, C.C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79-97.

Lee, S. J., Liu, Y. E., & Popović, Z. (2014). Learning Individual Behavior in an Educational Game: A Data-Driven Approach, *Proceedings of the 7th International Conference on Educational Data Mining*, London, UK.

Mory, E. H. (2003). Feedback research revisited. In D. H. Jonassen (Ed.*), Handbook of research on educational communications and technology* (pp. 745-783). Mahwah, NJ: Erlbaum.

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

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

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

Noweski, C., Scheer, A., Büttner, N., von Thienen, J., Erdmann, J., & Meinel, C. (2012). Towards a paradigm shift in education practice: Developing twenty-first century skills with design thinking. In H. Plattner, C. Meinel, & L. Leifer (Eds.), *Design thinking research* (pp. 71-94). Berlin Heidelberg: Springer. DOI: 10.1007/978-3-642-31991-4_5

Oppezzo, M.A. & Schwartz, D.L. (2013). A behavior change perspective on self-regulated learning with teachable agents. In R. Azevedo, & V. Alevan (Eds.), *International handbook of metacognition and learning* (pp. 485-500). New York: Springer. DOI: 10.1007/978-1-4419-5546-3_31

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994-1020.

Peacock S. B., Smith, C., Martin, T., Aghababyan, A., Popović, Z., Andersen, E., & Liu, Y. E. (2013). Learning fractions through splitting in an online game. Paper presented at the 2013 Annual Meeting of the American Educational Research Association, San Francisco, CA.

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

Razzouk, R., & Shute, V. (2012). What is design thinking and why is it important? *Review of Educational Research*, *82*(3), 330-348. DOI: 10.3102/0034654312457429

Ridley, D.S., Schutz, P.A., Glanz, R.S., & Weinstein, C.E. (1992). Self-regulated learning: The interactive influence of metacognitive awareness and goal-setting. *Journal of Experimental Education, 60*(4): 293-306.

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. DOI: 10.1016/j.learninstruc.2010.07.004

Roll, I., Baker, R. S., Aleven, V., & Koedinger, K. R. (2014). The effect of overuse and underuse of help resources in intelligent tutoring systems. *Journal of the Learning Sciences,* doi: 10.1080/10508406.2014.883977

Rotherham, A. J., & Willingham, D. (2009). 21st century. *Educational Leadership*, *67*(1), 16-21.

Ryan, R.M. & Deci, E.L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78.

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

Sadler, P.M., Coyle, H.P., & Schwartz, M. (2000). Engineering competitions in the middle school classroom: Key elements in developing effective design challenges. *Journal of the Learning Sciences, 9*(3), 299-327.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age.* Cambridge, MA: MIT Press. Open access title: https://mitpress.mit.edu/books/measuring-what-matters-most

Schwartz, D. L., Cheng, K. M., Salehi, S., & Wieman, C. (2016). Commentary: The half-empty question for socio-cognitive interventions. *Journal of Educational Psychology*, *108*(3), 397- 404. DOI: 10.1037/edu0000122

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

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA, US: MIT Press. Open access title: https://mitpress.mit.edu/books/stealth-assessment

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

Silk, E.M., Higashi, R., Shoop, R. & Schunn, C.D. (2010). Designing technology activities that teach mathematics. *The Technology Teacher*, 69(4), 21-27

Snow, E. L., Allen, L. K., Russell, D. G., & McNamara, D. S. (2014). Who's in control?: Categorizing nuanced patterns of behaviors within a game-based intelligent tutoring system. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining*, (pp. 185-192). London, UK.

Spörer, N., Brunstein, J. C., & Kieschke, U. (2009). Improving students' reading comprehension skills: Effects of strategy instruction and reciprocal teaching. *Learning and Instruction 19(*3), 272-286. DOI: 10.1016/j.learninstruc.2008.05.003

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

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. New York, NY,US: John Wiley & Sons.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568-2572. https://doi.org/10.1016/j.chb.2013.06.033

Wirth, K.R. & Perkins, D. (2008). Learning to learn. Recovered from: *http://www. macalester. edu/geology/wirth/CourseMaterials.html*

Zimmerman, B.Z. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*(2): 64-70.

Appendix A

Instructional Sequence Descriptions for Math, Social Studies, and Science Units

Though the specifics of the curriculum were not at test in this study, we provide descriptions of the instructional sequences to give readers a concrete sense of the design-thinking instruction and the differences between the treatments. The school used a mix of block and regular scheduling, so each day of instruction involved either a 45 or 90-minute period.

Math Design Lesson

This lesson was modified from a project used by the mathematics teachers in previous years. The project is to design a box that holds 18 small cylindrical candies. Students should design the box to be visually interesting and appealing, so customers would want to buy the candy, but they should also keep in mind cost of materials and waste if many boxes were to be manufactured. The mathematical content of the unit primarily focused on 2-dimensional nets and their correspondence to 3-dimensional objects, as well as concepts related to area, volume, and measurement. The seven days of instruction was largely the same for both treatments, with treatment differences on days 2 and 5.

Day 1. Students in both treatments started the activity by getting an overview of the design challenge and examining several actual candy boxes, which they could take apart. In small groups, they discussed features they noticed about the different boxes, such as shape and whether they packed the candies tightly or had empty space. Students then started learning about creating nets (described as 2D unfolded versions of 3D shapes) in preparation for designing their own boxes. They taped six small squares of paper together to create a cube. They then undid some of the tape until the cube was flattened into a two-dimensional shape. Students sketched the 2D net. Students then tried taping the six papers together in different configurations to determine which configurations could be successfully folded into a cube, and which could not. As a class, students discussed the properties of the configurations that could fold up into a cube. In pairs they were asked to think about additional criteria that might matter for choosing among different net designs if they were manufacturing a cube-shaped box, beyond whether it successfully made a cube. Ideas included how many nets could fit on a single large piece of cardstock and how many cuts each net required.

Day 2. Students continued to focus on nets more generally; they were instructed to try to make a new 3D shape. They could make any shape they wanted. When they had a 3D shape and associated net they were happy with, they created a second copy of the net. Then instruction varied by treatment.

Stakeholder-Feedback treatment (Feedback). Students were explicitly told that it is helpful to seek feedback about your ideas, so you can make them better. They paired off with another student in the class. One student gave her un-cut net to her partner, who cut it out and tried put it together into a 3D shape. The student interviewed the partner to find out what the he thought about the final shape, and also the process of putting together the net. For example, if students eyeballed the measurements and the sides were uneven, the partner might say that it was difficult to tape together because the sides did not match up well. Students were encouraged to seek suggestions for improvement. After both partners had exchanged feedback, the class was solicited for things they heard that could be useful for everyone. Then a new constraint was introduced that the shapes would have to be glued together, not taped. This necessitated the design of overlapping tabs that provided a surface to apply the glue. Afterwards, students modified or created a new design based on the feedback and to meet the new glue constraint.

Parallel-Explore treatment (Explore)*.* Students were explicitly told that it can be useful to look across several different ideas to find patterns, which might help them notice things they didn't think of before. Students got into small groups and put together all their 3D shape designs. They looked at the spread of designs and sorted them based on different criteria, such as volume and how many cuts would be required to make them. In addition to prompted criteria, student groups were encouraged to come up with their own criteria to sort by (e.g., sturdiness.) Then the additional constraint was introduced that the shapes would have to be glued not taped (requiring overlapping tabs). Groups sorted the shapes by how easy it would be to add tabs for glue. After this, students modified their design or created a new one to meet the new glue constraint, based on the compare and sort activity.

Day 3. Students in both treatments were introduced to the design cycle shown in Figure 5. As a class, they discussed each step of the design process in the context of planning a party. Students were then reminded that the candy box design challenge required their box to hold 18 candies. Highlighting the Explore phase of the design cycle, students manipulated wood 'candies' in pairs to explore different configurations that the 18 candies could be stacked in a box (e.g., one column of 18, 6 rows of 3, etc.). Then each table group was given a role (customer, factory manager, candy shop owner). As a class, students discussed things someone in each role might care about when it came to candy boxes (e.g., a storeowner might care about attractiveness and also shelf space; a factory manager might care about amount of cardstock used per box). Each group then got several examples of real candy boxes and evaluated them based on their role. They discussed what they found appealing and what problems they saw with the designs, which they then shared out with the class. Finally, students worked with a partner to brainstorm and prototype their own box design using graph paper, adhering to the criteria that it needed to hold the 18 wooden candies.

Day 4. All students continued prototyping their box and brainstormed a name, logo, and color scheme. Students were then asked to come up with a second box idea that was quite different from their first to generate different kinds of ideas.

Day 5. Instruction varied by treatment.

Feedback treatment. It was reiterated that seeking feedback can help improve ideas. The researchers acted in the role of candy shop owners. Students were told "Pick your favorite designs. You can ask us for feedback about what we think, and for suggestions for how they might be improved." Students had to come up to the researchers (candy shop owners) to request feedback. The researchers had several dimensions on which they could give feedback, including how it fit the candies, sturdiness/protectiveness, how it opened and closed, shape, visual design and uniqueness. Students then revised their designs.

Explore treatment. It was reiterated that examining different ideas to find patterns can help improve ideas. In small groups, students took all their candy box designs and put them in a line. They sorted them based on how packed the candies would be (from not fitting to having lots of extra room) and discussed which approaches they thought were best. They also sorted by how the boxes would open and close, how well they'd protect the candies, and by their own criteria, such as how well they'd stand out in a store, examining the different approaches across the different designs. Students then revised their designs or created a new design.

Days 6 & 7. All students finished their designs and made a final net on graph paper, as well as a final 3D box using cardstock. Students measured the dimensions of their box. They calculated the surface area and volume of the box, the volume of the candies, and the amount of excess space not used by candies. Students drew a picture of their box from a side angle, showing how the candies fit inside. Finally, they wrote a description of their box design and described why they made the design decisions they did.

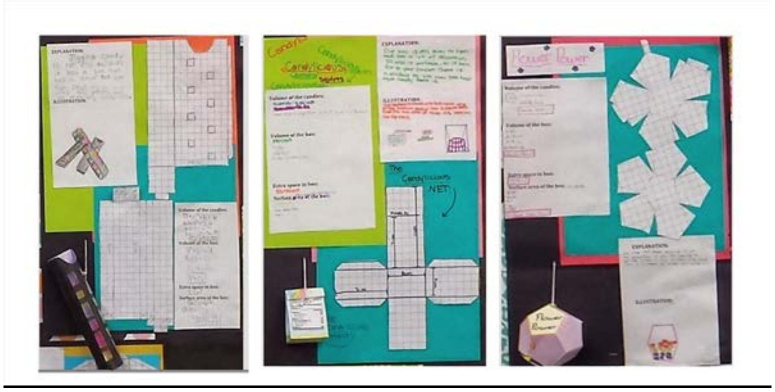


Figure A1. Three sample candy box designs.

Social Studies Design Lesson

The students' social studies design project was to make a method for achieving fair decisions. They learned that their method would be used in a final activity, and the decision achieved by their method would determine what would happen. The central topical goal was two-fold. The first was to revisit different governing structures across history (e.g., Roman Republic, Monarchies, Democratic Republic). The second goal was to help them understand the concept of a process for making good decisions. At this age, students are inclined to make their decisions, and then find a justification. Pushing them to think about the process of decision making in advance of a specific question is a demanding abstraction that ideally can help them understand different governing systems developed over the ages. The first day, a 45-minute class, introduced them to the design problem. The second day, a 90-minute class, reminded them about exploring the space of alternatives (Explore treatment) or seeking feedback from stakeholders (Feedback treatment), and then had them engage in a treatment-appropriate variation of a card game we created for decision making. On the third day, there was discussion and vote (majority wins) on the decision method the class would use. They then received the decision question of whether to stay in class to do extra school work versus go early to lunch.

Day 1. The lesson began for all students with a description of the design challenge: Design a process for making smart and fair decisions. We then reminded the students that we begin with the explore phase and why it is important in design. Next the students engaged in a couple of activities to help them understand the challenge. For instance, in one activity, they learned that the school bought iPads, but not enough for all the students. They then saw the following fictitious student proposals for how to distribute them.

- Jack: Children who have internet at home.
- Jose: Children who have behaved very well.
- Sarah: Children who know a lot about computers.
- Fred: Children who are boys.
- Sam: Children who are not very good at sports.
- Jen: Children who get good grades.
- Sally: Children who need help to get good grades.

It ended with the question of how should we decide? The instructors led a discussion to help the students understand that the question is not which proposal is the best, but rather, how do we create a process that we can use to decide which proposal is best. To help students make this switch and to connect with their prior social studies lesson, instructors asked students:

- How would a kingdom decide?
- How would the USA decide?
- How would the Ancient Greeks have decided?
- How would the Ancient Chinese have decided?
- How would your parents decide?

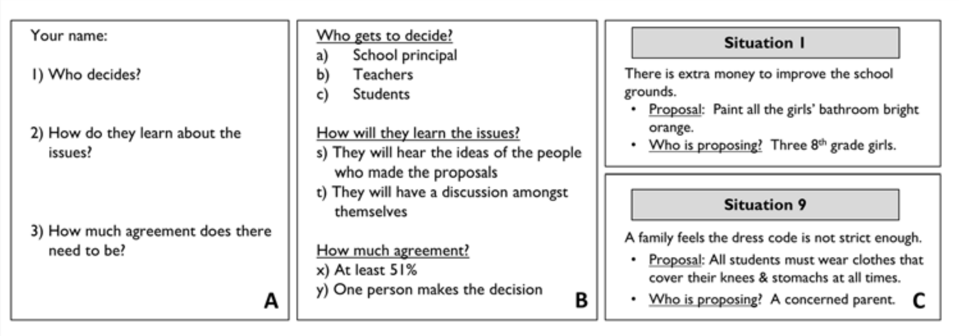
Following discussion, students wrote down how they would decide. Next, instructors led them through a consideration of some of the strengths of different systems:

- Majority vote: Good for picking between two proposals or people.
- 100% agreement: Good when you have to be sure (like a murder trial).
- Single ruler (e.g., monarch): Good when people do not have knowledge to make decisions.
- Representative democracy: Good when there are too many decisions for everyone.
- Direct democracy: Good for decisions that everyone cares about.

The day's activities ended with a task where the students worked in small groups to generate situations where each of these models for making decisions would work poorly. Instructors also provided a brief wrap up on the Explore phase of the design cycle.

Day 2. The day began for all students with a reminder of the design challenge to create a process for making fair decisions. This was followed by short presentation on the importance of the Generate phase of the design cycle and why prototyping one's initial ideas leads to better final solutions. This presentation included two historical examples, the development of a "better" parachute and a "better" way to pick tomatoes. The presentations differed by treatment to emphasize a specific design strategy. For example, the "better" parachute example either attributed the failure of the parachute to a lack of seeking feedback or a lack of considering alternatives.

All students were then introduced to the design activity of the day to create "decision cards," which were index cards on which kids filled out blanks for "WHO decides?", "HOW do they learn about the issue?", and "HOW MUCH agreement is needed?" See Figure A2.



Your name:

1) Who decides?

2) How do they learn about the issues?

3) How much agreement does there need to be?

Who gets to decide?
a) School principal
b) Teachers
c) Students

How will they learn the issues?
s) They will hear the ideas of the people who made the proposals
t) They will have a discussion amongst themselves

How much agreement?
x) At least 51%
y) One person makes the decision

Situation 1

There is extra money to improve the school grounds.
- Proposal: Paint all the girls' bathroom bright orange.
- Who is proposing? Three 8th grade girls.

Situation 9

A family feels the dress code is not strict enough.
- Proposal: All students must wear clothes that cover their knees & stomachs at all times.
- Who is proposing? A concerned parent.

Figure A2. A) Decision cards that students filled out to create a decision process. B) The abridged menu, used during early rounds, from which students could choose options. Later rounds expanded the menu to twice the number of options. In the final round, students were allowed complete freedom in creating their final, "best" solution for a fair decision process, filling out the fields for Who, How, and How much? however they wished. C) Examples of "Situation cards" against which students' decision cards were tested and discussed.

Feedback treatment. Students were asked to create a single decision card. Then, in small groups, they were to test their decision cards against a "situation card" chosen from a stack of possible situations. Each student was to gather feedback on the fairness of her decision card for the chosen situation, i.e. was the outcome a fair decision and what were perceived difficulties with the process.

The feedback process was iterated for two situations (rounds), before students were asked to generate a second decision card (from a menu of expanded options). Students' second decision cards were then tested in rounds 3 and 4, using two new situation cards and the same feedback and discussion process.

Explore treatment. Students were asked to create two different decision cards using the abridged menu of option shown in Figure A2.B. In small teams, students pooled their cards and were asked to sort them on the factor of "Who decides?" They tested the decision cards against a "situation card" chosen from a stack of possible situations, and were asked to discuss how the different groups of cards handled the situation. For example, did a particular group of cards provide more fair decisions? In the second round, students sorted on the factor of "How do they learn about the issue?" and repeated the test and discussion process with a new situation card. In round 3, students were asked to generate two more decision cards using an expanded menu of options, and conducted another round of sorting and discussion, using the factor "How much agreement?" and a new situation card. Finally, for round 4, students repeated the sort and discuss process, using the factor "How well did it handle the scenario?", and a new situation card.

For the final activity of the day, all students created their final, "best" decision card. These cards were unconstrained, children could fill out the fields of Who, How, and How much however they wished, and these cards were to be used the following day in a variety of "real life" situations.

Day 3. Students' final decision cards from each class were tallied the night before by researchers, and the list of options for each of the factors - "Who decides?", "How do they learn about the issue?", and "How much agreement is needed?" - were projected at the front of the class for discussion. These options were discussed in the context of the various types of situation cards, contrasting the "fairness" of the different decision processes for the different types of scenarios (e.g. public vs private situations, small vs. large groups of people affected, student vs. adult proposers, etc.). Then, children were each given an envelope containing a game chip of random color (red, blue, green, or yellow), a random playing card, as well as a random stack of play money, ranging in value from $12 - $1200. Next, four actual students' decision cards, representing a range of options, were projected at the front of the class and used to make decisions on a variety of situations. For example, one issue entailed early release from class. The proposals were a) Children with blue or yellow chips could leave early, b) Red and green chips could leave early, c) Children allowed to leave early are chosen by chance, and d) No one gets to leave early. There was also an issue involving taxation on the play money to pay for a hypothetical school bowling alley. The proposals were a) a fixed dollar amount of $20, b) a flat tax rate of 15% would be imposed, and c) a progressive tax rate would be imposed, with rates of 0%, 10% or 20%. Discussion on what was “fair” for this last issue, in particular, was quite lively.

Science Design Lesson

The science design project was to make a game that would teach about ecosystems. The central topical goal was to develop student understanding of what contributes to the stability of an ecosystem based on how energy flows through the various trophic levels (primary producers, primary consumers, secondary consumers). In the first three days of the unit, students played and made modifications to a few pre-existing games, which were based on aspects of ecosystem dynamics. On the third and fourth days, students designed their own ecosystem game, often using the first three games as inspiration.

Days 1 & 2. Instruction varied by treatment.

Feedback treatment. Feedback students began the unit by playing an ecosystem-themed version of a card game, War. The cards in the ecosystem deck each showed an organism that was a producer, primary consumer, or secondary consumer. Two players each drew an organism card from their deck for head-to-head comparison. The card from the higher trophic level would “munch” the lower card and that player would keep both cards. In the event of a tie (e.g., two producers were drawn), players would lay three more cards face down before drawing a fourth card to determine who gets the whole pile. After a few games, the students worked in pairs to modify the rules to make it into a “peace” game, which requires making the ecosystem as stable as possible, so nobody would win. They made their own rules, such as eliminating the 3-card draw upon a tie. Day 2 implemented the key treatment difference. Instructors re-introduced the design method of seeking stakeholder feedback, familiar from the mathematics and social studies units. Pairs swapped their new rules, and played one another’s games based on those rules. Acting as stakeholders, the pairs now wrote feedback about the game they had received. Feedback included constructive criticism on gameplay (e.g., missing rules, overly complex rules) and science content (e.g., unrealistic rules that did not correspond to real organism or ecosystem behavior). The original rule-writers received and processed the relevant feedback. Afterwards, there was a brief discussion of the importance of seeking constructive criticism from stakeholders.

Explore treatment. Explore students used the same deck of cards as the Feedback treatment to play a variation of musical chairs. Members of eight-student groups each picked one card from the deck. Then they walked around a table until the music stopped, at which point they matched up head-to-head with whomever happened to be sitting across from them at that moment. The higher trophic level card would “munch” the lower card and that player would move on to the next round. The rounds continued until there were no organisms left. They played this game several times, with each student keeping a data sheet indicating the initial spread of trophic levels in the group of eight and how many rounds the game lasted. Day 2 implemented the treatment difference. Instructors re-introduced the design method of parallel design and reminded students that they had played multiple games, each with a different spread of organisms and outcomes. They could use this source of variation to help figure out what was happening. Students met in new groups to compare their data from the day before. To encourage them to explore the space of outcomes, they sorted the data in a few different ways and discussed what they noticed. Almost everyone noticed that the more stable ecosystems were bottom heavy, with more organisms at the lower trophic levels. Afterward, instructors led a brief discussion of the importance of generating lots of ideas or data and looking through the results for patterns.

Days 3 & 4. All students played a new game, *Eco Chaos!,* a multi-player board game that modeled more aspects of ecosystems. The goal was to get everyone across the finish line. Rather than a single organism winning the game, the group of players wins the game by getting enough of their team members across. This models the stability of an entire ecosystem. Each group member picked a class of organism at one of three trophic levels (producer, primary consumer, or secondary consumer) and rolled dice to advance across the board to the finish line. Students started with fifteen energy points and expended one point for each square they moved. They gained more energy by passing (eating) another player at a lower trophic level, by reaching an “oasis” square on the board, or by landing on a “Chaos!” square. Players who landed on a Chaos square drew a *Chaos!* card, which could include good news (cloudless day gives +2 energy points to all producers) or bad (consumers get an illness, lose 2 energy points). Wasted/used energy went to a discard pile, illustrating the idea that most energy in an ecosystem is not passed on through eating.

After playing a few rounds, students received a day and a half to design their own game to teach about ecosystems. The students came up with a fun variety of games building from the earlier games or using game mechanics borrowed from games they knew, such as, *Battleship* or *Chutes and Ladders*. In all cases, students reflected on the principles of ecosystems their games represented. Students enjoyed designing their eco-system games, and they were surprisingly sophisticated in their thinking. For example, when students had to add an element that violated nature, so-to-speak, they were articulate about what rules they had violated for the sake of the game play.

The unit concluded with a brief lecture on how energy moves through an ecosystem, focusing on how 10% of the energy at one trophic level is used by the next trophic level. The lecture described how the 10% principle explains the stability of bottom-heavy ecosystems that have sufficient producers to support the consumers.

Appendix B

Descriptions of Statistical Analyses Conducted in Study

B1 – Missing Data Imputation

The main text (see *Participants*) outlines data collected from study participants. These include external standardized achievement scores (e.g., English-language arts, science, and math), as well as outcome measures derived from the three choice-based assessments (CBAs; Poster-Feedback, Photo-Explore, Apex-Explore): the strategy choices, in-game problem-solving performance, and post-test scores on learning. The group of students whose data is used throughout this analysis had recorded strategy choices in each of the three CBAs (n = 197). However, data such as standardized achievement and post-test scores were not available for all 197 students. The data loss is summarized in Table B1. To preserve statistical power, relevant missing data from the subset of 197 students were imputed.

Table B1. *Data loss statistics throughout the various study stages.*

Data Type		Variable	Number Missing (*n*=197)*
Student Achievement		Standardized ELA	44
		Standardized Science	54
		Standardized Math	108
Poster-Feedback	Strategy Use	Critical Feedback Choices	0
	Problem-Solving Performance	Poster Quality Gain	22
	Learning Outcome	Posttest Score	62
Photo-Explore	Strategy	Unique Camera Settings Tried	0
	Problem-Solving Performance	Aggregate Photo Quality	1
	Learning Outcome	Posttest Score	24
Apex-Explore	Strategy Use	Number of Requested Examples	0

*The superset of n = 197 students includes children for whom we have strategy choice data for each of the three CBAs. “Number missing” is defined as “number reported” subtracted from 197.

Missing values were drawn using multiple imputation by fully conditional specification (FCS MI; Van Buuren et al, 2006). Compared to other multiple imputation methods, which draw missing values from a pre-specified probability density function, FCS MI generates sequential imputations by prescribing an imputation model for each variable given the others (Carlin & Louis, 2008). This technique is particularly useful when the data are of different scales and when relationships between variables are unlikely to be modeled by more common parametric joint distribution functions (e.g., multivariate normal). Imputed datasets used for this article were drawn using the FCS MI routine of the SPSS (v25) software package.

Particular missing variables of interest included prior student achievement data (e.g., standardized ELA, science, and math scores) and CBA outcome measures (e.g., problem-solving performance scores and post-test learning scores). As such, the imputation model included each of these variables, as well as additional student-level data: academy (the school used three academies that determined which cohorts of teachers would teach which students), math tracking (the school tracked students by regular, advanced, and double advanced math), and science period, which determined students' teachers throughout the day. Together, these student-level tracking data defined intact classes of students. All variables except post-test scores were used as predictors in the imputation. Post-test scores were viewed as the final (temporal) outcome measure in the analysis, thus their use in the imputation model was solely as a dependent variable.

Ten complete imputed datasets were drawn for use in the analyses. Where possible, statistics were compiled from the pooled collection of the ten sets. In some cases, the functionality to pool from the imputations was unavailable in the statistical software. In these cases, we summarize the results of the ten imputations, for example, by showing the range of results across the imputations. In the test of achievement by treatment on choice behaviors, we present the results for the original data and the imputed data. In the remaining analyses, we only present results based on the multiple imputation.

B2 – Student Strategy Choices as a Function of Achievement and Treatment

We implement a multivariate, multilevel mixed-model (Goldstein, 2011) to predict the effects of treatment, average student achievement, and the interaction between them on students' strategy choices in three CBAs. We control for differences between individual students and behavior across classrooms by including students and classrooms as random effects in the model. The complete regression model is given by

$$y_{kij} = \beta_{1k} C_{kij} T_{ij} \ + \beta_{2k} C_{kij} S_{ij} \ + \beta_{3k} C_{kij} T_j S_j + u_{ij} + v_{ij} + \varepsilon_{kij} C_{kj} \qquad (1)$$

where *k* indexes over the three CBAs ($k = 1$: Poster-Feedback; $k = 2$: Photo-Explore; $k = 3$: Apex-Explore), *i* indexes over the 12 science classrooms, and *j* indexes over each student. The three strategy choice measures are given by y_{kij}.

The variable C_{kij} corresponds to the three CBAs of interest: Poster-Feedback, Photo-Explore, and Apex-Explore, respectively (i.e., C_{1ij} acts as an indicator variable for Poster-Feedback). The variable T_{ij} represents treatment. In this analysis, T_{ij} is treated as a covariate, with $T_{ij} = 0$ indicating the Explore treatment and $T_{ij} = 1$ indicating the Feedback treatment. (Note that this definition of T_{ij} holds throughout the analysis, and that T_{ij} does not reflect which treatment serves as the conceptual control for a given CBA. As a reminder for the reader, $T_{ij} = 0$ corresponds to the control condition for the Poster-Feedback and $T_{ij} = 1$ corresponds to the control condition for the Photo- and Apex-Explore CBAs.) The variable S_{ij} represents the average of the *z*-scored standardized ELA, science, and math test scores per student.

The model given in Eq. 1 estimates the best-fit parameters of interest: regression coefficients, random intercepts, and residual errors. Each β_{1k} represents the treatment contribution to the strategy choice measures. Similarly, each β_{2k} represents the effect of achievement. We implement a version of Eq. 1 using only the β_{1k} and β_{2k} regression coefficients (Model A). The β_{3k} corresponds to the regression parameters illustrating the effect of the treatment by achievement interaction. Inclusion of this parameter is motivated by the observation that for each CBA the treatment appears to have a greater impact on the lower-achieving students. We implement a second version of Eq. 1 using β_{1k}, β_{2k}, and β_{3k} to estimate the contribution of this interaction on the strategy measures (Model B). Parameters u_{ij} and v_{ij} represent the random intercepts controlling for classrooms and students within classrooms, respectively. Finally, the three ε_{kij} correspond to the residual errors associated with each CBA. These random intercepts and residual error are included in both Models A and B. Best-fit results using ten imputed datasets for models A and B are in Table B2**.**

Table B2. *Best-fit parameter estimates of Eq. 1 not including the interaction (Model A) and including the interaction (Model B) of condition and achievement, using pooled results from ten imputed datasets.*

Model	CBA	β_{1k}[a]	β_{2k}[a]	β_{3k}[a]	BIC[b]
A	Poster-Feedback: $k = 1$	0.12 ± 0.14	0.23 ± 0.11*	---	1690.77
	Photo-Explore: $k = 2$	-0.03 ± 0.14	0.23 ± 0.10*	---	
	Apex-Explore: $k = 3$	-0.15 ± 0.14	0.12 ± 0.11	---	
B	Poster-Feedback: $k = 1$	0.11 ± 0.13	0.43 ± 0.16**	-0.33 ± 0.19†	1691.84
	Photo-Explore: $k = 2$	-0.02 ± 0.13	0.06 ± 0.14	0.32 ± 0.20	
	Apex-Explore: $k = 3$	-0.13 ± 0.13	-0.17 ± 0.15	0.53 ± 0.19**	

[a] Parameter estimates are quoted with standard errors; ** p < .01; * p < .05; † p < 0.1.
[b] The average Bayesian Information Criterion (BIC) for each model as determined using each of the 10 imputed datasets is included as a measure for model selection.

B3 – Mediation Analyses

Mediation analyses explored the direct and indirect effects of strategy choices, problem-solving performance, and post-test learning measures. In particular, the mediation analyses determined if strategy choices flowed through problem-solving performance to affect post-test learning, while controlling for achievement and treatment. Univariate analyses for Photo-Explore and Poster-Feedback used the PROCESS macro for SPSS (Bolin 2014, Hayes 2013)**.** In both cases, choices and performance were serial moderators, and post-test score was the final dependent variable. The regression models, excluding the intercepts, are depicted in Figures 7 and 8 in the manuscript**.** At the time of writing, PROCESS did not have the functionality to pool over multiple imputed datasets. Therefore, the mediation analyses were run on each complete imputed dataset separately. Results are summarized in Tables B3 and B4.

Table B3. *Parameter estimates from mediation analysis on Poster-Feedback posttest.*

Input Node	Output Node	Average Parameter Estimate[a]	p_{min}[b]	p_{max}[b]
Average Achievement	Critical Feedback Choices	0.31 ± 0.08**	0.000	0.005
Treatment	Critical Feedback Choices	0.35 ± 0.01*	0.006	0.019
Average Achievement	Poster Quality Gain	0.35 ± 0.09**	0.000	0.001
Treatment	Poster Quality Gain	0.04 ± 0.14	0.481	0.985
Critical Feedback Choices	Poster Quality Gain	0.14 ± 0.07†	0.016	0.188
Average Achievement	Post-test Score	0.27 ± 0.06**	0.000	0.002
Treatment	Post-test Score	0.03 ± 0.09	0.082	0.096
Critical Feedback Choices	Post-test Score	0.01 ± 0.05	0.309	0.944
Poster Quality Gain	Post-test Score	0.68 ± 0.05**	0.000	0.000

[a] Reported average parameter estimates are the average values over the ten imputed datasets; reported standard errors are the averages of the standard errors over the ten imputed datasets. Overall significance is determined using the average of the ten imputed datasets; ** p < .01; * p < .05; † p < 0.1.

[b] For each imputation, the significance of a nonzero effect (*p*) was determined; p_{min} and p_{max} represent the range of *p*-values computed over the ten imputed datasets.

Table B4. *Parameter estimates from mediation analysis on Photo-Explore posttest.*

Input Node	Output Node	Average Parameter Estimate[a]	p_{min}[b]	p_{max}[b]
Average Achievement	Unique Camera Settings Tried	0.30 ± 0.09**	0.000	0.003
Treatment	Unique Camera Settings Tried	0.04 ± 0.14	0.640	0.912
Average Achievement	Aggregate Photo Quality	0.10 ± 0.08	0.157	0.545
Treatment	Aggregate Photo Quality	-0.04 ± 0.13	0.668	0.888
Unique Camera Settings Tried	Aggregate Photo Quality	0.36 ± 0.07**	0.000	0.000
Average Achievement	Post-test Score	0.16 ± 0.09	0.002	0.734
Treatment	Post-test Score	0.01 ± 0.14	0.167	0.966
Unique Camera Settings Tried	Post-test Score	0.17 ± 0.10†	0.000	0.402
Aggregate Photo Quality	Post-test Score	0.10 ± 0.08	0.043	0.153

[a] Reported average parameter estimates are the average values over the ten imputed datasets; reported standard errors are the averages of the standard errors over the ten imputed datasets. Overall significance is determined using the average of the ten imputed datasets; ** p < .01; * p < .05; † p < 0.1.

[b] For each imputation, the significance of a nonzero effect (*p*) was determined; p_{min} and p_{max} represent the range of *p*-values computed over the ten imputed datasets.

Appendix B References

Bolin, J. H. (2014). Hayes, A. F. (2013). Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, NY: The Guilford Press. *Journal of Educational Measurement*, *51*(3), 335-337.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Goldstein, H. (2011). *Multilevel Statistical Models* (Vol. 922). John Wiley & Sons.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049-1064.