The Design Implementation Framework: Guiding Principles for the Redesign of a Reading Comprehension Intelligent Tutoring System

Kathryn S. McCarthy[1], Micah Watanabe[2], and Danielle S. McNamara[2]

[1] Georgia State University

[2] Arizona State University

The Design Implementation Framework: Guiding Principles for the Redesign of a Reading Comprehension Intelligent Tutoring System

Abstract

The Design Implementation Framework, or DIF, is a design approach that evaluates learner and user experience at multiple points in the development of intelligent tutoring systems. In this chapter, we explore how DIF was used to make system modifications to iSTART, a game-based intelligent tutoring system for reading comprehension. Using DIF as a guide, we conducted internal testing, focus groups, and usability walk-throughs to develop iSTART-3, the latest iteration of iSTART. In addition to these evaluations, DIF highlights the need for experimental evaluation. With this in mind, we describe an experimental evaluation of iSTART-3 as compared to its predecessor, iSTART-ME2. Analyses revealed an interesting tension between system usability and user preference that has important implications for instructional designers.

Keywords: Design implementation framework, Intelligent tutoring system, reading comprehension, Game-based tutoring

The Design Implementation Framework: Guiding Principles for the Redesign of a Reading Comprehension Intelligent Tutoring System

Kathryn S. McCarthy, Micah Watanabe, & Danielle S. McNamara

The Design Implementation Framework, or DIF, is a design approach that evaluates learner and user experience at multiple points in the development of intelligent tutoring systems. In this chapter, we explore how DIF was used to make system modifications to iSTART, a game-based intelligent tutoring system for reading comprehension. Using DIF as a guide, we conducted internal testing, focus groups, and usability walk-throughs to develop iSTART-3, the latest iteration of iSTART. In addition to these evaluations, DIF highlights the need for experimental evaluation. With this in mind, we describe an experimental evaluation of iSTART-3 as compared to its predecessor, iSTART-ME2. Analyses revealed an interesting tension between system usability and user preference that has important implications for instructional designers.

1. Introduction

Intelligent tutoring systems, or ITSs, provide the opportunity for individualized computer-based instruction, evaluation, and feedback at scale. ITSs are effective learning tools—students who engage with ITSs show learning gains similar to one-on-one human tutoring or small group instruction (Ma et al., 2014; VanLehn, 2011). Advances in technology and pedagogy mean that ITSs are constantly evolving to be “better, faster, and cheaper” (Craig et al., 2018). Thus, iterative modifications are a critical aspect of ITS development. These modifications should not only be theory-driven and empirically-validated, but also practically-valuable for a variety of stakeholders (Craig, 2018; Roscoe et al., 2017). While meaningful educational gains are the key outcome for ITSs, other aspects of the ITS experience are also important to acknowledge. However, little work has been published on usability and experience in intelligent tutoring systems (Chughtai et al., 2015; Lin et al., 2014). The Design Implementation Framework (DIF; Stone et al., 2018) was developed to address this gap in ITS design and user experience.

In this chapter, we outline DIF and describe key aspects of the framework in the context of foundational design approaches, such as ADDIE. We then present a case study in which we used the DIF in the redesign of the reading comprehension ITS iSTART. Guided by principles of DIF, we conducted participatory research that included teachers and students throughout the development process and an iterative development, implementation, and evaluation cycle. The result of this effort is an improved system that is not only more accessible but also more engaging and effective.

2. Design Implementation Framework

DIF (Stone et al., 2018) is an emerging framework for instructional designers that connects research, design, and implementation processes. DIF is a cycle composed of five phases: (a) Defining and Evaluating the Problem, (b) Ideation, (c) Design and User Experience, (d) Experimental Evaluations, and (e) Feedback and Implementation (Table 1).

DIF is founded upon existing methods of instructional design—such as the Analysis, Design, Development, Implementation, Evaluation model (ADDIE; Molenda, 2003) and Design-based Implementation Research (DBIR; Fishman et al., 2013)—but was developed with specific consideration of the affordances and constraints of intelligent tutoring systems. Further, DIF is a design approach that takes into consideration a variety of end users. For example, teachers play an important role in the success of educational technology in the classroom, yet instructors are often ignored as both facilitators and end-users (Stone et al., 2018). DIF is part of a larger effort by researchers in education, cognitive psychology, and the learning sciences that encourages participatory design and educators-as-partners in the development and refinement of educational technologies (Luckin & Cukurova, 2019).

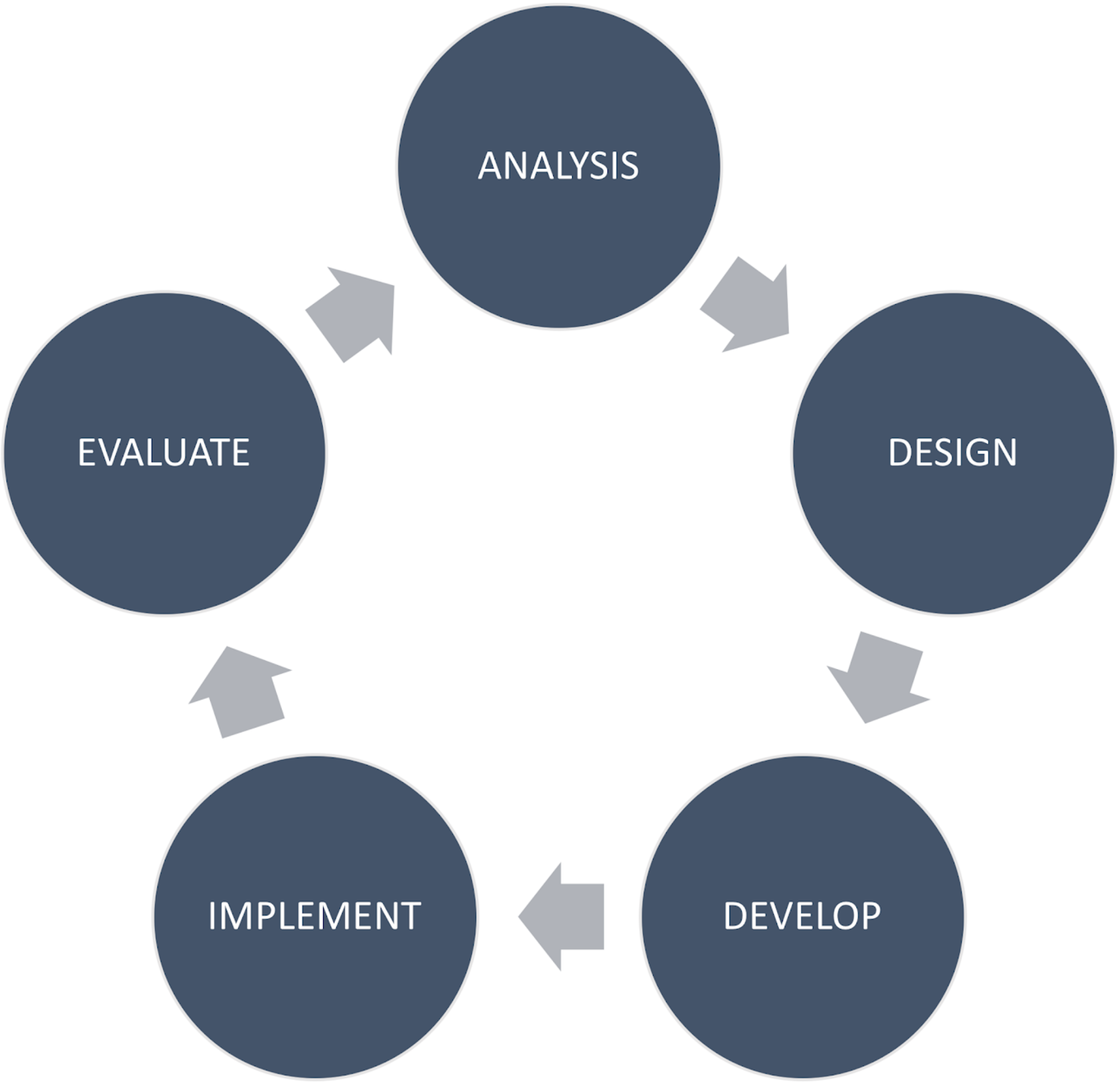
Table 1

Phases of the Design Implementation Framework

Note: Adapted from Stone et al. (2018).

DIF Phase	Description
Defining and Evaluating the Problem	Identifying one or more central research questions or problems emerging from the developers’ instructional or theoretical goals.
Ideation	A creative and collaborative brainstorming process to generate a variety of plans for implementation or further investigation.
Design and User Experience	Usability and user experience research methods to test and refine the designs (e.g., sketches, mockups, paper prototyping, and wire framing).
Experimental Evaluations	Evaluation of new hard-coded interface and fully-functional system features to assess impact on learning, motivation, and other outcomes of interest via laboratory or school-based experiments.
Feedback and Implementation	Deployment of the technology in authentic learning settings (e.g., classrooms).

DIF differs from design sequences such as ADDIE and DBIR, both in terms of specific phases of development and the structure of those phases. For example, a quick Google search for ADDIE yields diagrams that generally fall into one of two layouts. The first (Figure 1) indicates a unidirectional loop, starting with analysis and ending with evaluation. The second (Figure 1) indicates a loop including the first four aspects, with evaluation at the center, presumably to reflect its impact at each stage of development.


ANALYSIS
DESIGN
DEVELOP
IMPLEMENT
EVALUATE

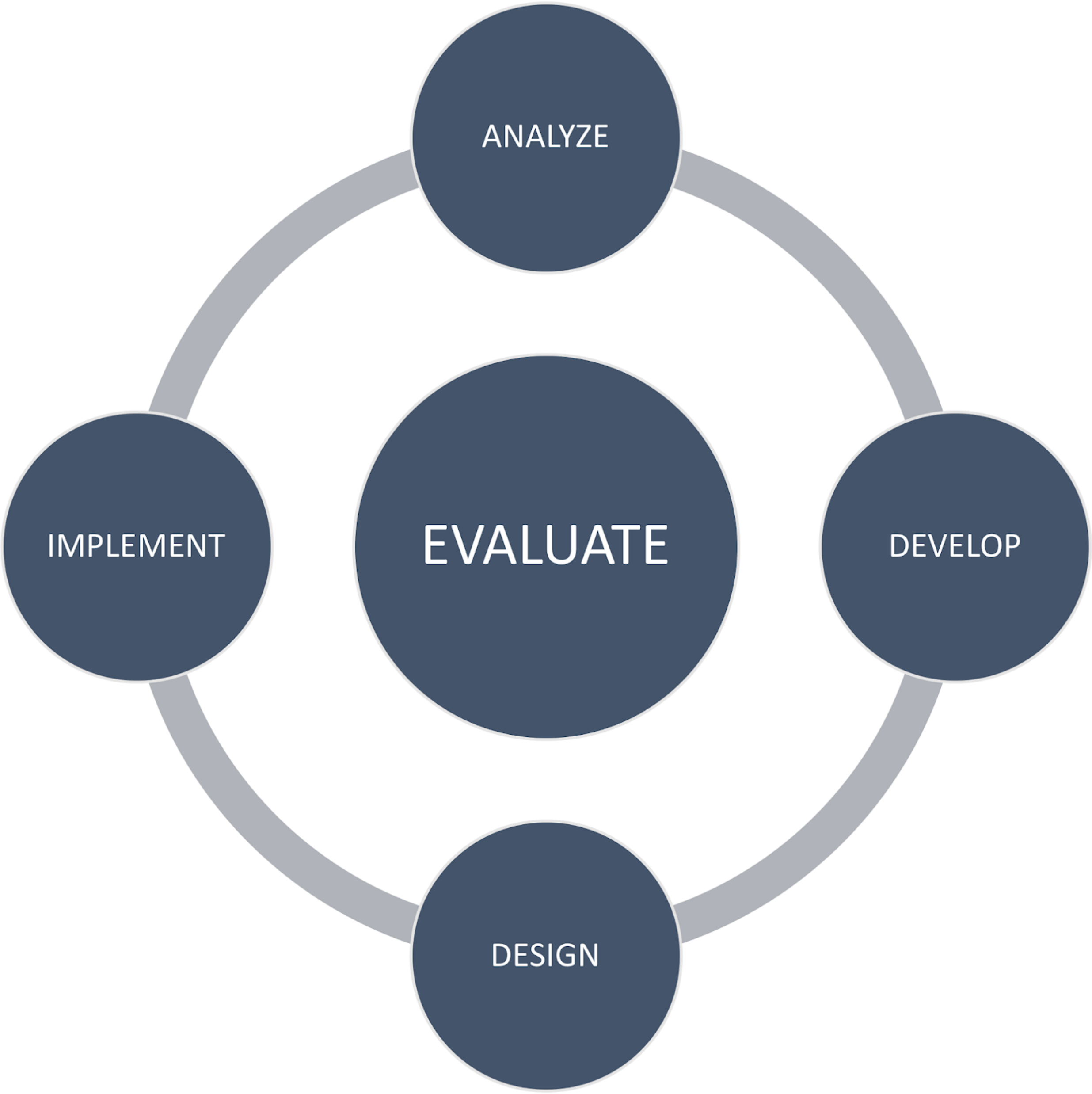


Figure 1

Common Diagrams of ADDIE Model

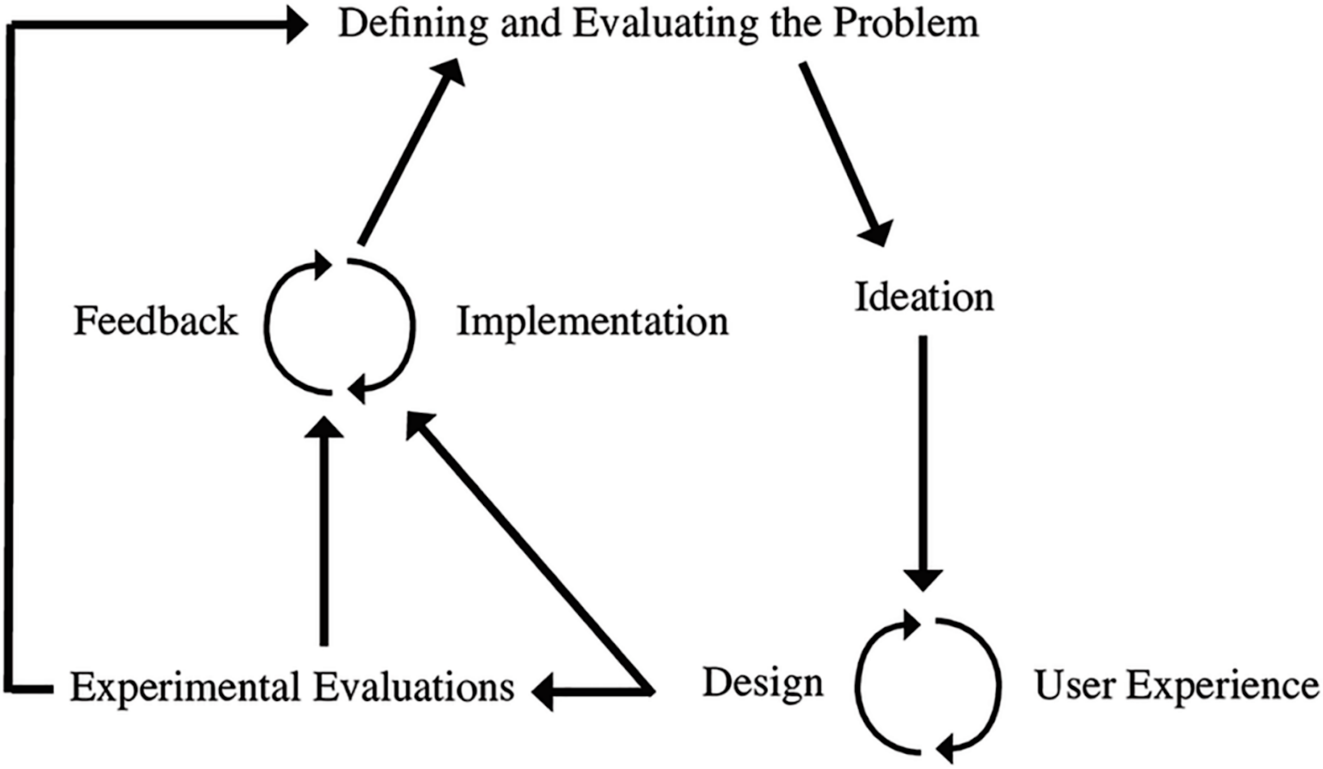


Figure 2

The Design Implementation Framework

Note: From Stone et al. (2018).

In contrast, DIF is conceptualized with feedback and evaluation at various points along iterative refinement (Figure 2). This complexity reflects the diverse and potentially conflicting outcomes relevant to successful ITSs. ITSs must first and foremost support learning, but other aspects of design can help or hinder learning gains. Craig and colleagues (2004) demonstrated that boredom during ITS use is negatively correlated with learning. A pedagogically-motivated system modification might demonstrate increased learning in short lab trials; if the users find the system boring, they may not learn as much or engage with it long enough for tutoring to have substantial long-term effects (Jackson & McNamara, 2011, 2013). Alternatively, designers might introduce new features to increase and maintain interest. However, if these features are distracting, they can cause learners to engage in unproductive, off-task behaviors that do not support learning (Rowe et al., 2009). Good ITS design requires finding a “sweet spot” of a system that is easy to use, enjoyable, and efficacious. Thus, a key element of the DIF cycle is experimental evaluations in addition to user feedback.

For example, we recently tested the effect of two metacognitive prompts (McCarthy et al., 2018). We were motivated by research in both reading comprehension and ITS development showing that increasing metacognitive awareness can improve learning (Azevedo et al., 2016; Snow, 2002). We developed and designed two types of metacognitive prompts to help students better monitor their performance. We implemented these two prompts into a beta-test classroom within our ITS, iSTART (described below), and compared the effects of the prompts (independently and in combination) to the version of iSTART without these features. In a sample of more than 100 students, we found that the addition of these prompts did not lead to learning gains above and beyond the original iSTART practice environment architecture. Critically, the prompts also lead to decreased performance during system use, especially for less-skilled readers (McCarthy et al., 2018). Based on these findings, these metacognitive prompts were not implemented as default options into iSTART.

Experiments allow researchers to provide strong evidence that the changes made to the system have meaningful impacts on a variety of dimensions (e.g., motivation, perceptions, time-on-task, learning) prior to full implementation. By conducting evaluations incrementally and across various dimensions, we can continuously monitor the balance between differential outcomes in order to improve the system in ways that are both user-friendly and impactful.

3. Case Study: iSTART

In the remainder of the chapter, we describe how DIF was used to guide additional iterations of redesign of our ITS, iSTART. Interactive Strategy Training for Active Reading and Thinking (iSTART) is an intelligent tutoring system that uses video lessons, guided instruction, and game-based practice to improve students' reading comprehension skills through self-explanation training. Self-explanation, or the act of explaining a text to yourself during reading, has been shown to be an effective learning strategy across a variety of domains (Bisra et al., 2018; Chi, 2000). Further, instruction on how to produce high quality self-explanations during reading improves students' comprehension of complex scientific texts (McNamara, 2004, 2017). iSTART leverages natural language processing to provide automated self-explanation instruction and feedback to improve reading comprehension skill.

In iSTART, students are introduced to five self-explanation strategies: comprehension monitoring, paraphrasing, predicting, bridging, and elaborating. These strategies have been shown to improve comprehension across a variety of age ranges and skill levels (e.g., Cain & Oakhill, 1999, 2006, 2011; McNamara et al., 2006; Palincsar & Brown, 1984). The strategies are introduced in brief video lessons and then students are introduced to a practice environment. In Coached Practice, students practice reading texts and writing their own self-explanations. Natural language processing-based algorithms guide both summative and formative feedback. A summative score from 0-3 is presented on the overall quality (i.e., *poor*, *fair*, *good*, or *great*) of the self-explanation. Formative feedback is provided by a pedagogical agent who offers targeted feedback messages to help students revise their self-explanations.

Students can also play generative games or identification games. In the generative games, students earn points for writing higher quality self-explanations. In the identification games, students read example self-explanations and earn points for identifying which strategy is being demonstrated. These points can then be used to purchase more game play or to purchase accessories for the player's avatar.

The iSTART system has undergone several iterations. The original iSTART (McNamara et al., 2004) was a computer-based version of the in-person intervention, Self-Explanation Reading Training (SERT; McNamara, 2004). iSTART included video lessons and guided practice with feedback. iSTART-2 (Levinstein et al., 2007) improved the self-explanation scoring and feedback algorithm and used classroom-based data to make improvements to the existing modules. iSTART-ME (Motivationally-Enhanced; Jackson et al., 2009; Jackson & McNamara, 2013) introduced the game-based practice environment. While the games themselves do not improve comprehension, they improve students' motivation, which, in turn, mediates their learning gains and continued training (Jackson & McNamara, 2011, 2013). iSTART-ME was updated to iSTART-ME2 in order to incorporate a teacher interface. This required reprogramming the system using a combination of Java and Flash (Snow et al., 2016). Each of these versions were built based on design research between our research team, teacher-partners, and student participants. Building on this tradition of redesign and reevaluation, we set out to use DIF to develop the next generation of iSTART reading comprehension training.

3.1. Defining and Evaluating the Problem

The first phase of DIF requires designers to define and evaluate the problem. To identify issues that were most relevant to our end users, we iteratively worked with teachers and students to identify weaknesses in the existing version of the system, iSTART, and the barriers that might prevent teachers from using iSTART effectively. We conducted focus groups and worked closely with teacher-partners who were implementing iSTART (in this case, iSTART-ME2) into their classrooms. We surveyed students from these classes as well as users from our lab-based studies. These experiences revealed three aspects of iSTART in need of redesign.

The first major concern we heard from our teacher-partners was that the existing iSTART system could only be run on desktop or laptop computers. As mobile technology has become more affordable, tablets have become more prevalent in classrooms (Burke & Hughes, 2018). Many of our teacher-partners had ready access to tablets, whereas they would need to reserve space in computer labs in order to use iSTART during the school day. We also took into consideration that students from lower socio-economic status homes tend to rely on smartphones and tablets for connectivity (Li et al., 2015; Tsetsi & Rains, 2017). Thus, some students have restricted access to engaging in additional practice at home. The solution to this problem was relatively direct. By recoding iSTART from Flash to HTML5, we were able to offer responsive design (e.g., mobile compatibility). Although this was a straightforward change, the actual recoding of the system required extensive effort on the part of the programmers and designers (as well as federal funding from the Office of Naval Research and the Institute of Education Sciences).

A second problem identified was that teachers and students found the overall graphics and design of iSTART to be outdated. This was not too surprising as the system was developed in the early 2000s and only superficial aesthetic changes had been made in the interim. Users also noted that they disliked the cold, text-to-speech narration used in the training videos and suggested using real voices. This feedback was not new—teachers and students had previously complained about the automated speech engines (Levinstein et al., 2007). At that time, the automated voices were not replaced because key aspects of iSTART were still under development. We had relied on automated voices so that these revisions could be made relatively quickly without needing to re-record and re-edit the content as iterative changes were being implemented and evaluated. Since that time, the content of the lessons has concretized. Thus, the ability to use recorded voices was now more practical.

The third and final problem identified during the problem definition phase was that the teacher interface introduced in iSTART-ME2 was difficult to navigate. One benefit of iSTART's text-general algorithm is that instructors can import their own texts to tailor lessons to specific classes or students. However, our teacher-partners found importing and assigning texts to be cumbersome. Teachers who struggle to make the ITS work quickly and easily are not likely to integrate tutoring into their class time. Even the best-designed learner tool may cease to have an impact if the instructor does not integrate its use into the classroom. The teachers also noted that the student progress and performance pages provided useful data, but that it would be beneficial to have these data aggregated in meaningful ways that could help them to diagnose issues more quickly.

3.2. Modernizing the Look of iSTART

With the problems defined, we moved to the next phases of DIF: ideation and design. We began ideation by informally examining trends in website and game design. The research team met to discuss which designs were most appropriate for iSTART. As we moved into the design phase, the research team met weekly to discuss and develop potential designs. Individual team members presented mockups of the interface or particular game elements, and the team iteratively compared designs and offered feedback. By having multiple potential designs, the team was able to compare and contrast these options. As we progressed, these discussions led to a single design that "mixed and matched" the best aspects of the different designs. These design discussions led us to replace the cartoon icons with photographs to represent the various strategies and redesigned the games to have more realistic looks and feel. The fonts, tabs, and buttons were modernized accordingly (Figure 3). We updated the training videos to be consistent with the overall interface design. Once we were satisfied with our mockup designs, we shared them with our teacher-partners. They reacted positively toward the new interface and expressed that their students would like the new design and videos. We used this feedback to move more confidently into implementation.

To address concerns over the narration, we elected to record human narrators. The team considered theoretical and practical constraints when determining the sounds of our narrator(s). Ultimately, we decided on two male voices and a female voice. We decided on multiple narrators for several reasons. We first considered theoretical implications—research indicates that the gender of an instructor (or pedagogical agent) can impact students' perceptions and learning (Baylor & Kim, 2004; Elias & Loomis, 2004) and that these effects are driven by whether the instructors' gender is the same or different from that of the learner (Krämer et al., 2016). Including different "instructors" of differing genders allowed us to reduce potential bias. The second reason was more practical. Having multiple voices facilitates adding in other voices if content needs to be edited or added, and, if new videos are added in the future, without the need to entirely re-record old versions to maintain consistency.

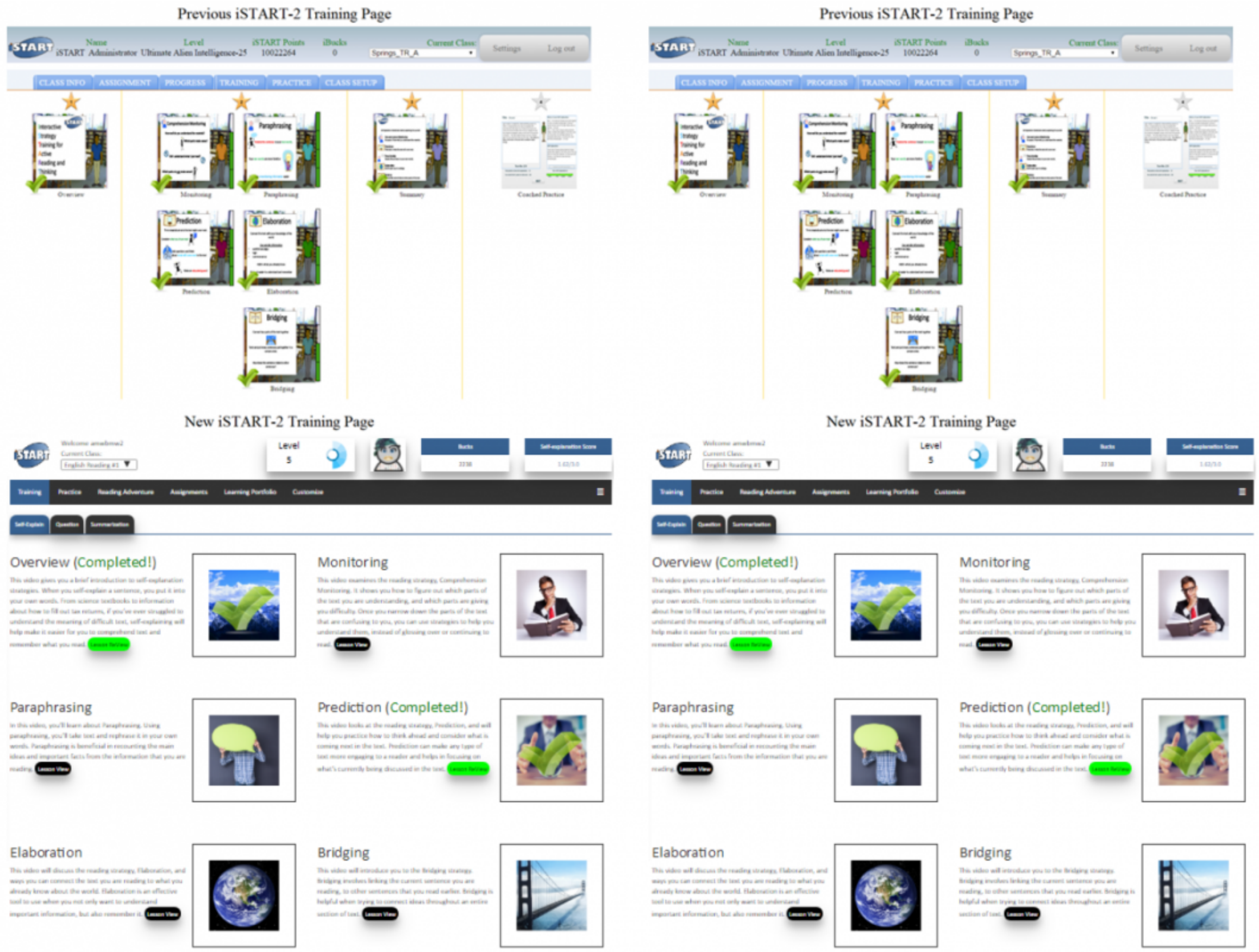
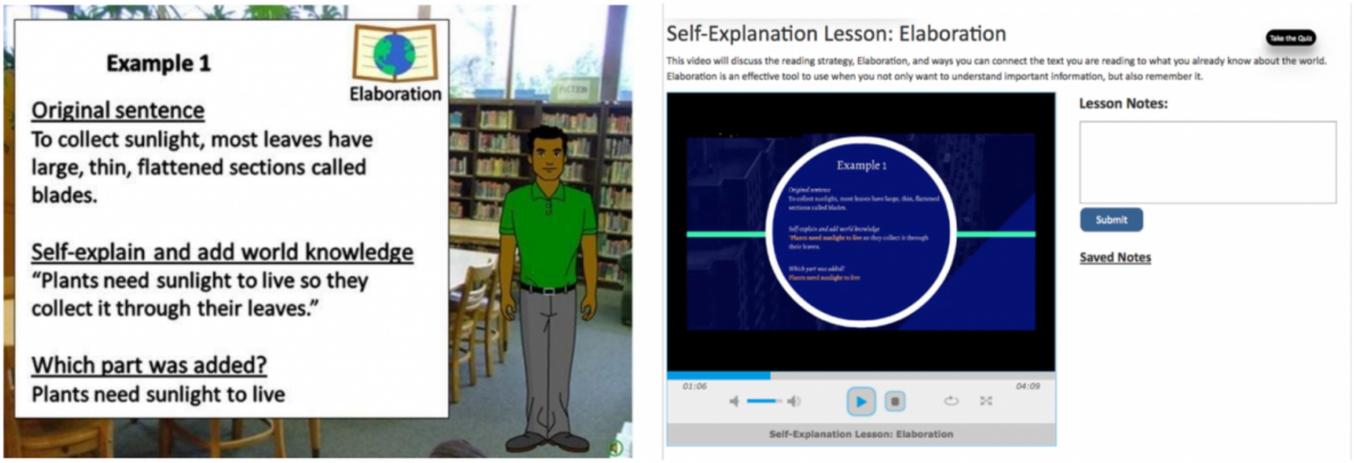


Figure 3a

Previous iSTART Training Menu and the Redesigned HTML5 iSTART-3 Training Menu



Original sentence
To collect sunlight, most leaves have large, thin, flattened sections called blades.

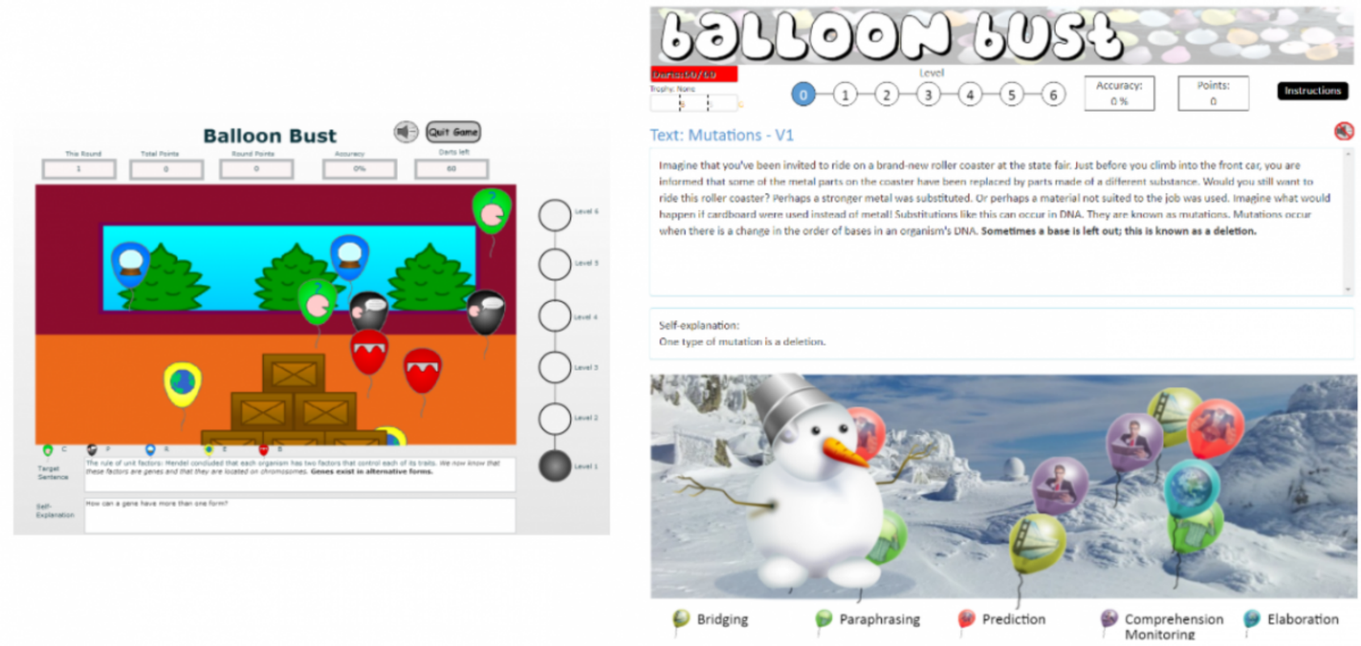
Self-explain and add world knowledge
"Plants need sunlight to live so they collect it through their leaves."

Which part was added?
Plants need sunlight to live

Self-Explanation Lesson: Elaboration

This video will discuss the reading strategy, Elaboration, and ways you can connect the text you are reading to what you already know about the world. Elaboration is an effective tool to use when you not only want to understand important information, but also remember it.

Figure 3b
Previous iSTART Video and the Redesigned HTML5 iSTART-3 Video Interface



The rule of unit factors: Mendel concluded that each organism has two factors that control each of its traits. *We now know that these factors are genes and that they are located on chromosomes.* **Genes exist in alternative forms.**

How can a gene have more than one form?

Imagine that you've been invited to ride on a brand-new roller coaster at the state fair. Just before you climb into the front car, you are informed that some of the metal parts on the coaster have been replaced by parts made of a different substance. Would you still want to ride this roller coaster? Perhaps a stronger metal was substituted. Or perhaps a material not suited to the job was used. Imagine what would happen if cardboard were used instead of metal! Substitutions like this can occur in DNA. They are known as mutations. Mutations occur when there is a change in the order of bases in an organism's DNA. **Sometimes a base is left out; this is known as a deletion.**

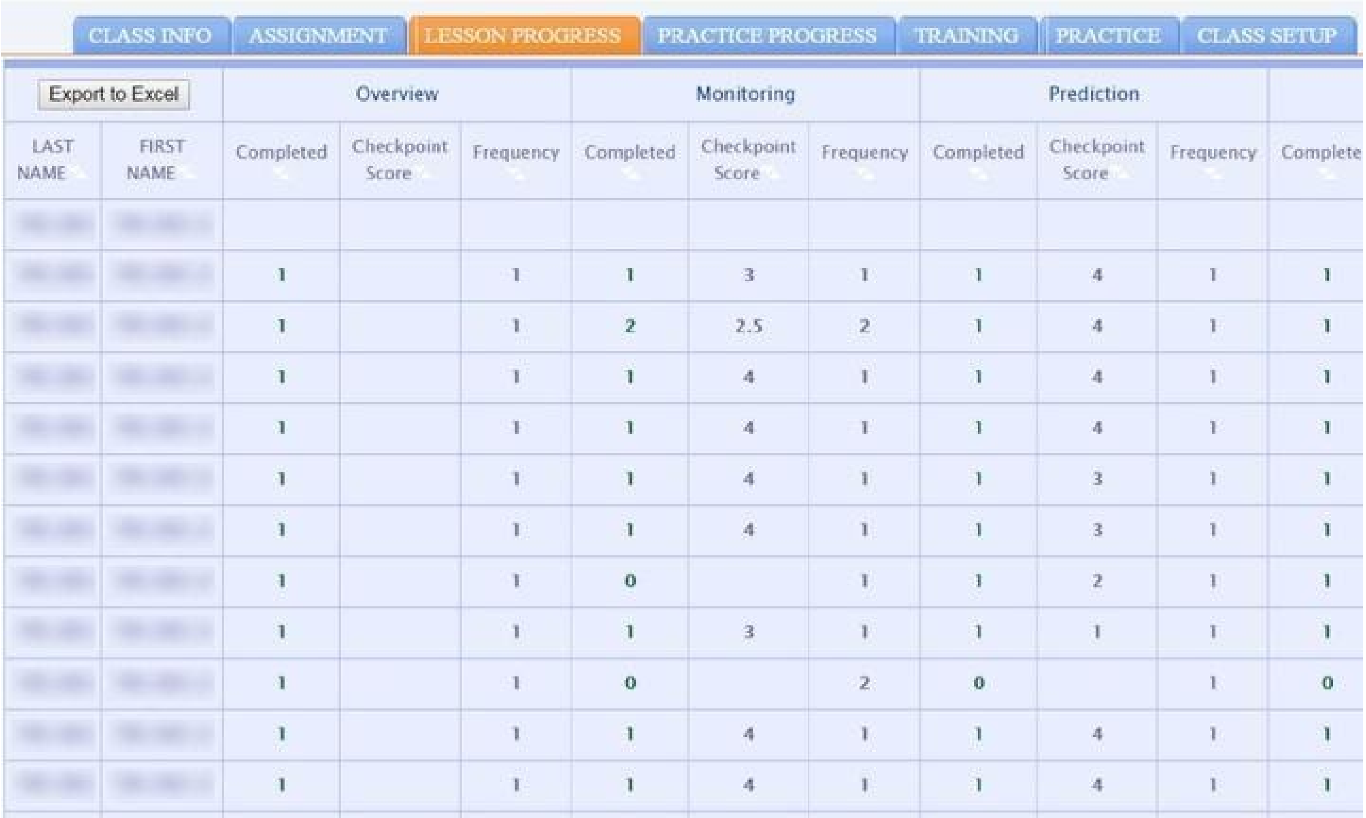
One type of mutation is a deletion.

Figure 3c
Previous iSTART Practice Game and the Redesigned HTML5 iSTART-3 Practice Game

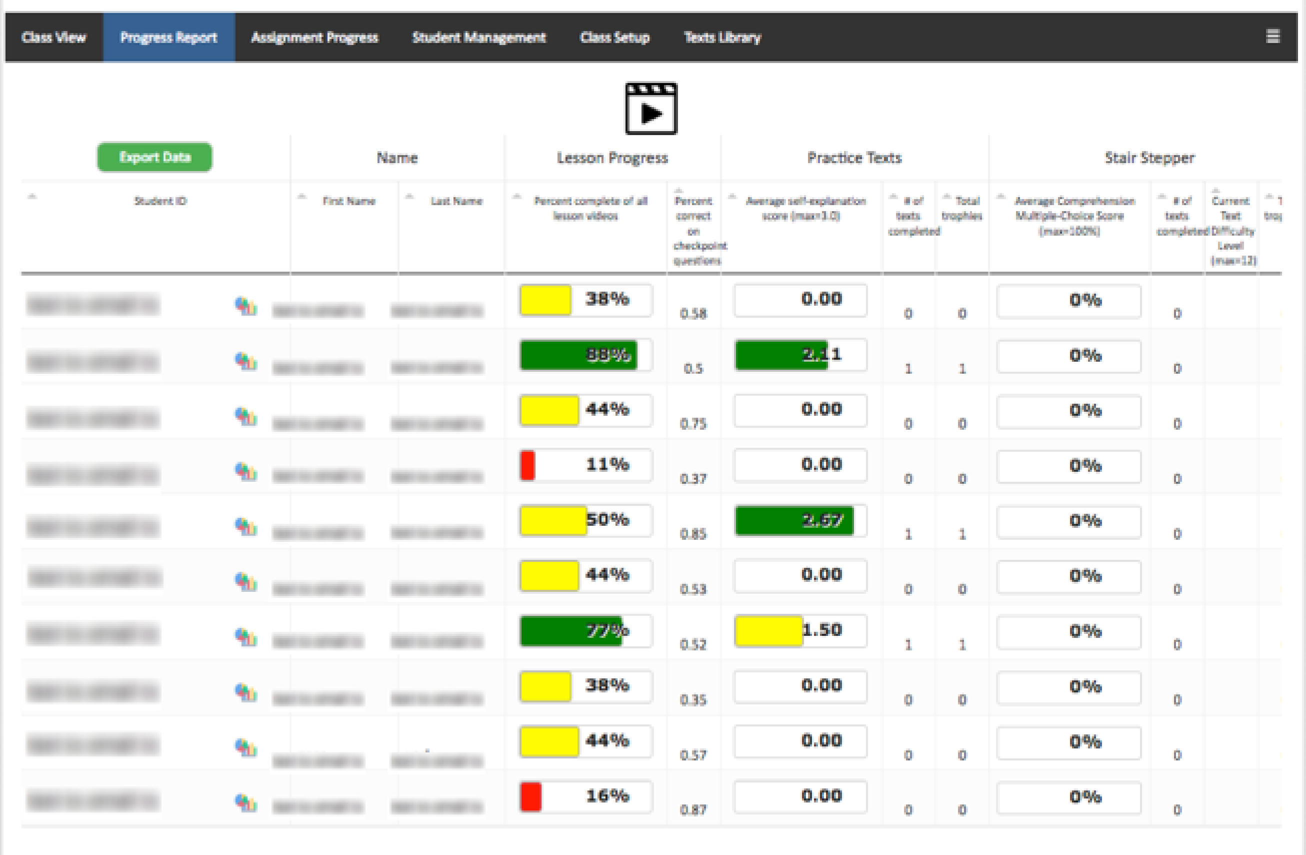
3.3. Improving the Teacher Interface

In parallel with the modernization of the system, we also began ideation for redesign of the teacher interface. The teacher interface includes two broad categories of information. The first is a calendar-based screen on which instructors can assign texts and modify deadlines. The second is a dashboard on which instructors can view student progress in terms of overall completion of videos and assignments as well as in terms of aggregate and individual self-explanation scores.

We conducted several focus groups and interviews with teachers who were using iSTART as well as teachers using its sister ITS, The Writing Pal (Roscoe et al., 2014), with the intent of developing these interfaces in parallel, with slight modifications for the specific needs of each system. These interviews helped us to define the specific aspects of the interface in need of redesign as well as to allow the teachers to join us in ideation. In order to gather user experience data, we constructed prototypes of the interface using the Marvel prototyping app. Prototyping apps and programs allow designers to generate interactive mockups in order to rapidly complete multiple cycles of design and experience prior to investing time and effort into hard coding the system. We adapted a cognitive walkthrough methodology (Lewis et al., 1990; Wharton et al., 1994) to collect user experience data. Cognitive walkthrough is a usability inspection method in which evaluators (e.g., developers, research participants) are asked to go through the system as if they were a user in order to identify weaknesses in design and functionality of a system. In most cases, evaluators are given a series of tasks that a user might need to complete. Evaluators talk-aloud about their process (e.g., Ericsson & Simon, 1998; Pressley & Afflerbach, 1995) as they complete these tasks, and their system behaviors are recorded and analyzed. We conducted two rounds of cognitive walkthroughs and redesign. This cycle helped us to simplify navigation and to better understand how users were interpreting the student performance data. We used what we learned from those experiences, as well as advances in dashboard design (e.g., Few, 2006), to drive the changes specific to the iSTART teacher interface. For example, in iSTART-ME2, lesson progress was displayed numerically. In iSTART-3, students' progress is represented through color-coded progress bars (Figure 4). Teachers can also view class-level data on a particular assignment or drill down to see individual student scores presented in a simple line graph.



		Overview			Monitoring			Prediction			
LAST NAME	FIRST NAME	Completed	Checkpoint Score	Frequency	Completed	Checkpoint Score	Frequency	Completed	Checkpoint Score	Frequency	Complete
[redacted]	[redacted]										
[redacted]	[redacted]	1		1	1	3	1	1	4	1	1
[redacted]	[redacted]	1		1	2	2.5	2	1	4	1	1
[redacted]	[redacted]	1		1	1	4	1	1	4	1	1
[redacted]	[redacted]	1		1	1	4	1	1	4	1	1
[redacted]	[redacted]	1		1	1	4	1	1	3	1	1
[redacted]	[redacted]	1		1	1	4	1	1	3	1	1
[redacted]	[redacted]	1		1	0		1	1	2	1	1
[redacted]	[redacted]	1		1	1	3	1	1	1	1	1
[redacted]	[redacted]	1		1	0		2	0		1	0
[redacted]	[redacted]	1		1	1	4	1	1	4	1	1
[redacted]	[redacted]	1		1	1	4	1	1	4	1	1



	Name		Lesson Progress		Practice Texts			Stair Stepper			
Student ID	First Name	Last Name	Percent complete of all lesson videos	Percent correct on checkpoint questions	Average self-explanation score (max=3.0)	# of texts completed	Total trophies	Average Comprehension Multiple-Choice Score (max=100%)	# of texts completed	Current Text Difficulty Level (max=12)	T tro
[redacted]	[redacted]	[redacted]	38%	0.58	0.00	0	0	0%	0		
[redacted]	[redacted]	[redacted]	88%	0.5	2.11	1	1	0%	0		
[redacted]	[redacted]	[redacted]	44%	0.75	0.00	0	0	0%	0		
[redacted]	[redacted]	[redacted]	11%	0.37	0.00	0	0	0%	0		
[redacted]	[redacted]	[redacted]	50%	0.85	2.67	1	1	0%	0		
[redacted]	[redacted]	[redacted]	44%	0.53	0.00	0	0	0%	0		
[redacted]	[redacted]	[redacted]	77%	0.52	1.50	1	1	0%	0		
[redacted]	[redacted]	[redacted]	38%	0.35	0.00	0	0	0%	0		
[redacted]	[redacted]	[redacted]	44%	0.57	0.00	0	0	0%	0		
[redacted]	[redacted]	[redacted]	16%	0.87	0.00	0	0	0%	0		

Figure 4

Teacher Interface from iSTART-ME2 and the Redesigned HTML5 iSTART-3

As we neared full implementation, we realized that these new visualizations could also benefit student users. Consistent with the DIF, we took the opportunity to redefine our problem to include the need for clearer data visualization for both teacher and student. Thus, the student progress screen was recoded with an identical colored-coded progress bar design and the ability to open these aggregate scores into game-level or text-level metrics.

3.4. Implementation and Feedback

After these rounds of design and user experience with the prototype, we hard-coded iSTART to reflect these design changes. After this implementation, we collected feedback for this phase by conducting a final cognitive walkthrough with the hard-coded system. We asked undergraduates (n = 5) to complete a series of target tasks that teachers and students would need to complete. The intent of this walkthrough was to emulate the needs of a teacher implementing iSTART in the classroom. The teacher would need to be able to quickly add new texts and assignments and to view student progress, but they would also need to be familiar with navigating the larger system to help guide students and troubleshoot when necessary. We used screen capture software to record system behavior (e.g., assigning a text to a class, playing a game, finding their average score on that game in the progress screen). One limitation in prompting thinking aloud is that it can potentially disrupt the natural cognitive processes that occur or encourage participants to engage in processes that would not have occurred without verbalizing (Branch, 2000; Nisbett & Wilson, 1977). Rather than asking participants to think-aloud, we conducted retrospective interviews to gain additional insights into user experience. In retrospective interviews, users are asked after the task to talk about what they did during the task. Although this approach has some limitations including potential bias or simple memory errors, we elected to use this less invasive method. Retrospective interviews also allow the researchers to ask additional follow-up questions to clarify or expand upon the user's responses. Data from the retrospective interviews indicated that the system was relatively easy to navigate, but that some of the tabs were labeled in ways that were unclear, leading to some confusion. For example, the majority of students had difficulty finding their average self-explanation score with the tab labeled "Texts Scores." In design, we had given this tab this shorter name so that the length of each tab was consistent. However, users preferred clarity of a tab's function over consistency of design. The users suggested in their interviews that the label "Self-Explanation Scores" would be clearer. The researchers involved in these walkthroughs brought these findings back to the research team and modifications were recommended to our programmers.

3.5. Experimental Evaluation

Experimental evaluation is a unique and critical aspect of DIF. Experiments, unlike other research methods, allow us to draw causal conclusions. True experiments (as opposed to quasi-experiments) require that users are randomly assigned to either the treatment or a comparison (i.e., "control") group. In experiments, the goal is to hold all other variables constant, so that the only difference across groups is the variable of interest (or the manipulation). For example, in the aforementioned study by Baylor and Kim (2004), the researchers used the exact same audio for their pedagogical agents and changed only the agents' appearance (e.g., race). By holding the audio constant, the researchers were able to more confidently conclude that learners' perceptions were based solely on appearance rather than other characteristics or behaviors. The ability to make direct comparisons is important for interpretive feedback. For example, a user rating of 3.75/5 might seem like an excellent score on its own. However, a comparison is necessary to contextualize any particular score. That is, if the previous version of the system was rated on average as 4/5, then a score of 3.75 may not be considered as high as it appears on the surface. Experimental evaluations can be a bit more time and resource intensive, but they can provide instructional designers greater confidence in the efficacy of their tools.

Guided by DIF, we conducted an experimental study to evaluate user perceptions of iSTART-3—the version of the system that includes responsive design, modernized aesthetics, and clearer dashboards. In some of our previous work, we conducted large-scale (n > 100), long-term experiments (e.g., more than 10 hours of system use; semester long implementations). While these approaches are certainly valuable and allow the opportunity to explore interactions with individual differences (see Jackson & McNamara, 2011; McCarthy et al., 2018), we also encourage designers to consider employing smaller-scale experimental designs with convenience samples. Such evaluations afford strong empirical evidence of efficacy without being cost or resource prohibitive. Indeed, this sort of repeated testing at increasingly larger scales is well-aligned with the focus of the DIF.

These cognitive walkthroughs had primarily focused on teachers as the end users. Within this cycle of system modifications, our next step was to focus on evaluating iSTART from the perspective of students as the end users. For this round of evaluation, we conducted a smaller scale study with a convenience sample of undergraduates. Undergraduates (n = 54) interacted with iSTART for about 3 hours. Three hours was enough time to complete the video lessons, as well as to have time to play a variety of practice games. The undergraduates were randomly assigned to work with iSTART in the new responsive design (iSTART-3) version or the previous version (iSTART-ME2). After interacting with the system, they responded to the questions presented in Table 2 using a 1 to 5 Likert scale. The control condition, iSTART-ME2, allowed us to directly compare how our design changes compared to the previous iteration.

Table 2
Average Likert Scale Ratings as a Function of iSTART Version

	iSTART-ME2 (*n* = 29)	**iSTART-3 (*n* = 25)**		
	M (*SD*)	*M* (*SD*)	*t*(52)	*p*
Training Videos				
I enjoyed the overall look and feel of the training videos.	2.90 (1.24)	3.52 (.87)	2.11	0.040
The narration used in the videos was easy to understand.	3.03 (1.40)	4.08 (.76)	3.33	0.002
I felt like I learned the material during today's session.	3.10 (1.11)	3.80 (.87)	2.54	0.014
Practice Games				
I enjoyed the overall look and feel of the practice games.	3.11 (1.32)	3.19 (1.30)	0.24	0.812
The games were enjoyable to play.	3.21 (1.34)	3.27 (1.08)	0.17	0.870
Overall Interface				
I enjoyed the overall look and feel of the iSTART interface	2.96 (1.29)	3.04 (.91)	0.24	0.810
When I wanted to know how well I was doing in iSTART, the information was easy to interpret	3.57 (1.23)	4.00 (1.10)	2.25	0.029

As shown in Table 2, there was little difference across versions in students' perceptions of the overall environment or the practice games. However, participants preferred the new look of the training videos and found the narration easier to understand. Participants who used the new version of iSTART also found it easier to interpret the data that was presented about their performance. These results suggest that our redesign of iSTART addressed end-user feedback about the system.

We also had students complete a usability survey, adapted from the System Usability Scale (Brooke, 1996). This 10-item measure can be administered and scored quickly and the survey items are written to be system general. That is, the SUS items do not need to be modified from tool to tool. Perhaps due to its ease of use, the SUS has been used thousands of times and has been demonstrated to be a robust tool (Bangor et al., 2008). Thus, the SUS is a low-cost, relatively high-impact tool for instructional designers.

Students responded to the 10 items about the usability of the system on a Likert scale from 1-7. Although the test can be administered on paper, we used the survey system, Qualtrics, to collect the self-reported SUS. The students' usability rating for iSTART-ME2 (*M* = 34.9, *SD* = 8.68) was marginally higher than the ratings from iSTART-3 (*M* = 30.3, *SD* = 10.3), *t*(49) = -1.75, *p* = .09. One potential explanation for this lack of difference is that iSTART-3 was essentially in its infancy. As such, we discovered bugs in the system that were less about design and more about growing pains of the system. For example, one participant noted that the lesson video suffered from an excessive lag time in responsiveness. Additionally, our experimenter observational notes indicate that some students were logged out of the system during practice and needed to log back in, which would be disruptive and understandably frustrating. In sum, our findings indicated that the new iSTART-3 interface showed significant improvements in aesthetics and interpretability of performance data as compared to iSTART-ME2, but that the new system was not more user-friendly and, if anything, was slightly less usable than its predecessor. Our findings from this experimental evaluation highlight the tension between different critical outcomes in ITS design and redevelopment. Thus, designers need to carefully examine how modifications influence a variety of factors related to system use. We are using these data to address potential bugs, but also to modify the system to be more usable, while monitoring that these changes do not have detrimental effects on learning.

Our next steps, guided by DIF, are to test iSTART-3 in authentic classrooms and collect feedback from both teachers and students. This level of evaluation will complete one full “cycle” of the Design Implementation Framework, but will provide the data necessary to guide problem definition in the next DIF cycle for iSTART.

4. Conclusions and Lessons Learned

In this chapter, we introduced the Design Implementation Framework and demonstrated how DIF guided improvements in the intelligent tutoring system (ITS) for reading comprehension, iSTART. The development of iSTART has been an iterative process that has resulted in several versions (Levinstein et al., 2007; McNamara et al., 2004; Snow et al., 2016) that reflect the state-of-the-art at the time they were created. By leveraging DIF, we have been able to integrate new technologies, such as mobile compatibility, while maintaining a system that is effective in terms of learning gains and that meets the needs and ever-changing demands of its end users. Data from cognitive walkthroughs and experimental evaluations showed positive effects of our redesign efforts. More specifically, iSTART-3 improved students’ perceptions of the ease of use and enjoyment of the training modules (which were modified), but did not affect students’ perceptions of the games (which were updated, but not modified). It is expected that iSTART will require further updates to meet the expectations of users and maximize the availability of iSTART as the standards for and capabilities of educational technology evolve.

This implementation of the DIF gave us a means of improving our ITS, iSTART, but it also gave us valuable insight into the framework itself. Although other frameworks (e.g., ADDIE, DBIR) do not preclude rapid cycling, DIF’s emphasis on feedback cycles within the larger design cycle encouraged us to continually test our ideas and modifications at each phase of development. The explicit inclusion of experimental evaluations also allowed us to uncover the inconsistency across system usability and user preference. One limitation to the present work is that we explored only self-reported preference and usability. To more fully understand different components of learner and user experience, future testing will be conducted to collect behavioral data as well as target learning outcomes.

Perhaps the most important lesson learned from our team, through the two decades of development with iSTART, is that usability and experience must be gauged across a series of iterative cycles of design, feedback, and evaluation. As DIF highlights, evaluation can take a variety of forms and should occur at multiple phases of design. DIF encourages user-centered design at all stages of the ITS life cycle. Notably, our users ranged from members of the research team, to lab-based participants, to classroom students, and to teachers. We encourage developing instructional designers to consider a variety of methods of evaluation, such as cognitive walkthroughs, short-term experimental comparisons, and longitudinal studies. DIF’s emphasis on multiple types of evaluation should encourage instructional designers to consider multiple end users as well as the many different types of outcomes that are relevant to high-quality intelligent tutoring systems. DIF has served us well in development (and redevelopment) of iSTART and its sister system, The Writing Pal (see Stone et al., 2018). We anticipate that this approach to design will be beneficial for additional ITSs and other educational technologies. However, conducting more research with the framework will be critical before generalizations can be made.

Those who are interested in developing educational technologies, and more specifically ITSs, should be driven by considerations of learning processes and performance gains. That is, the foundations of any quality system should be built upon sound learning theory. This focus on effective instruction, practice, and feedback inherently defines the development of systems that help students learn. However, instructional designers must also recognize the need to focus on usability and user experience. Solely examining learning gains or whether a student enjoys engaging with a system is not enough. Both learning and motivation are important considerations in the development and implementation of educational technologies. Equally crucial is the consideration of how educational technologies differentially impact different types of learners. The ultimate objective in the use of automated tutoring systems is to adapt to the needs of the users. As such, examining the effects of individual differences and adapting to those differences should remain a key priority.

References

Azevedo, R., Martin, S. A., Taub, M., Mudrick, N. V., Millar, G. C., & Grafsgaard, J. F. (2016). Are pedagogical agents' external regulation effective in fostering learning with intelligent tutoring systems? *Intelligent Tutoring System: 13th international Conference, ITS 2016, Zagreb, Croatia, June 7–10, 2016. Proceedings* (pp. 197–207). Springer. https://edtechbooks.org/-nxo

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*(6), 574–594. https://edtechbooks.org/-Sfk

Baylor, A. L., & Kim, Y. (2004). Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. *Intelligent tutoring systems: 7th international conference, ITS 2004, Maceió, Alagoas, Brazil, August 30–September 3, 2004, Proceedings*, (pp. 592–603). Springer-Verlag. https://edtechbooks.org/-wMiJ

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review, 30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research, 22*(4), 371–392. https://edtechbooks.org/-hLh

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry, 189*(194), 4–7.

Burke, A., & Hughes, J. (2018). A shifting landscape: Using tablets to support learning in students with diverse abilities. *Technology, Pedagogy and Education, 27*(2), 183–198. https://edtechbooks.org/-fuKT

Cain, K., & Oakhill, J. (1999). Inference making ability and its relation to comprehension failure. *Reading and Writing, 11*(5-6), 489–503. https://edtechbooks.org/-yWDi

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology, 76*(4), 683–69. https://edtechbooks.org/-DNbB

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities, 44*(5), 431–443. https://edtechbooks.org/-UoAo

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology, 5*, 161–238.

Chughtai, R., Zhang, S., & Craig, S. D. (2015). Usability evaluation of intelligent tutoring system: ITS from a usability perspective. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), (pp. 367–371). https://edtechbooks.org/-uyY

Craig, S. D. (2018). *Tutoring and intelligent tutoring systems.* Nova Publishers.

Craig, S. D., Graesser, A. C., & Perez, R. S. (2018). Advances from the Office of Naval Research STEM grand challenge: Expanding the boundaries of intelligent tutoring systems. *International Journal of STEM Education, 5*(1), 5–11. https://edtechbooks.org/-wCF

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29*(3), 241–250. https://edtechbooks.org/-EzEt

Elias, S. M., & Loomis, R. J. (2004). The effect of instructor gender and race/ethnicity on gaining compliance in the classroom. *Journal of Applied Social Psychology, 34*(5), 937–958. https://edtechbooks.org/-mNpG

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178–186. https://edtechbooks.org/-dQyD

Few, S. (2006). *Information dashboard design: The effective visual communication of data.* O' Reilly Media.

Fishman, B. J., Penuel, W. R., Allen, A-R., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *National Society for the Study of Education, 112*(2), 136–156.

Jackson, G. T., Boonthum, C., & McNamara, D. S. (2009). iSTART-ME: Situating extended learning within a game-based environment. In H. C. Lane, A. Ogan, & V. Shute (Eds.), *Proceedings of the Workshop on Intelligent Educational Games at the 14th Annual Conference on Artificial Intelligence in Education* (pp. 59–68). AIED.

Jackson, G. T., & McNamara, D. S. (2011). Motivational impacts of a game-based intelligent tutoring system. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 519–524). AAAI Press.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105*, 1036–1049. https://edtechbooks.org/-ugPU

Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., & Gratch, J. (2016). Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers & Education, 99*, 1–13. https://edtechbooks.org/-ihj

Levinstein, I.B., Boonthum, C., Pillarisetti, S.P., Bell, C., & McNamara, D.S. (2007). iSTART 2: Improvements for efficiency and effectiveness. *Behavior Research Methods, 39*, 224–232. https://edtechbooks.org/-tsd

Lewis, C., Polson, P. G., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. CHI '90: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 235–242). ACM. https://edtechbooks.org/-APTs

Li, J., Snow, C., & White, C. (2015). Teen culture, technology and literacy instruction: Urban adolescent students' perspectives. *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie, 41*(3). https://edtechbooks.org/-CByK

Lin, H. C. K., Chen, N. S., Sun, R. T., & Tsai, I. H. (2014). Usability of affective interfaces for a digital arts tutoring system. *Behaviour & Information Technology, 33*(2), 105–116. https://edtechbooks.org/-wXYH

Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology, 50*(6), 1–15. https://edtechbooks.org/-gjTY

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://edtechbooks.org/-ugd

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education, 28*(3), 1–19. https://edtechbooks.org/-bmF

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*(1), 1–30. https://edtechbooks.org/-WcLi

McNamara, D. S. (2017). Self-Explanation and Reading Strategy Training (SERT) Improves low-knowledge students' science course performance. *Discourse Processes, 54*(7), 479–492. https://edtechbooks.org/-jpa

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers, 36*(2), 222–233. https://edtechbooks.org/-TunU

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34*(2), 147–171. https://edtechbooks.org/-WIJ

Molenda, M. (2003). In search of the elusive ADDIE model. *Performance Improvement, 42*(5), 34–37. https://edtechbooks.org/-sPD

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://edtechbooks.org/-nkmo

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117–175. https://edtechbooks.org/-kaT

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Routledge. https://doi.org/10.2307/358808

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59. https://edtechbooks.org/-zkG

Roscoe, R. D., Craig, S. D., & Douglas, I. (Eds.). (2017). *End-user considerations in educational technology design.* IGI Global. https://edtechbooks.org/-JBZd

Rowe, J. P., McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2009). Off-task behavior in narrative-centered learning environments. *Proceedings of the 2009 conference on artificial intelligence in education: Building learning systems that care: From knowledge representation to affective modelling* (pp. 99–106). IOS Press.

Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension.* RAND.

Snow, E. L., Jacovina, M. E., Jackson, G. T., & McNamara, D. S. (2016). iSTART-2: A reading comprehension and strategy instruction tutor. In D. S. McNamara & S. A. Crossley (Eds.), *Adaptive educational technologies for literacy instruction* (pp.104–121). Taylor & Francis, Routledge. https://edtechbooks.org/-CEn

Stone, M., Kent, K. M., Roscoe, R. D., Corley, K. M., Allen, L. K., & McNamara, D. S. (2018). The design implementation framework: Iterative design from the lab to the classroom. In R. Roscoe, S. D. Craig, & I. Douglas (Eds.), *End-user considerations in educational technology design* (pp. 76–98). IGI Global. https://edtechbooks.org/-CUPb

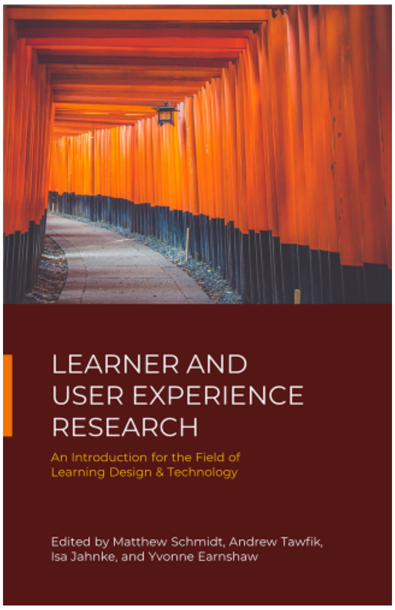
Tsetsi, E., & Rains, S. A. (2017). Smartphone Internet access and use: Extending the digital divide and usage gap. *Mobile Media & Communication, 5*(3), 239–255. https://edtechbooks.org/-EMmf

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://edtechbooks.org/-kYgV

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. Mack (Eds.), *Usability inspection methods* (pp. 105–140). Wiley.

Acknowledgements

This project was funded by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130124, R305A120707, R305A180261, R305A180144, R305A190050, and the Office of Naval Research, through Grant N00014140343, N000141712300, to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the Office of Naval Research. The authors acknowledge the extensive contributions by the lead designers and programmers: Amy Johnson, Jianmin Dai, and Haojun Pei. We also thank Cecile Perret, Laura Allen, Renu Balyan, Kathleen Corley, Matt Jacovina, Aaron Likens, Tricia Guerrero, Kevin Kent, Melissa Stone, Joseph Aubele, Carson Flood, Ashleigh Horowitz, Gary Ma, Amber Poteet, and our teacher-partners for their contributions to iSTART.


LEARNER AND USER EXPERIENCE RESEARCH
An Introduction for the Field of Learning Design & Technology
Edited by Matthew Schmidt, Andrew Tawfik, Isa Jahnke, and Yvonne Earnshaw

McCarthy, K. S. , Watanabe, M., & McNamara, D.S. (2020). The Design Implementation Framework: Guiding Principles for the Redesign of a Reading Comprehension Intelligent Tutoring System. In M. Schmidt, A. A. Tawfik, I. Jahnke, & Y. Earnshaw (Eds.), *Learner and User Experience Research: An Introduction for the Field of Learning Design & Technology*. EdTech Books. https://edtechbooks.org/ux/9_the_design_impleme

CC BY: This work is released under a CC BY license, which means that you are free to do with it as you please as long as you properly attribute it.