

# Designing a Prototype Tablet-Based Learning-Oriented Assessment for Middle School English Learners: An Evidence-Centered Design Approach

ETS RR–18-46

Carol A. Chapelle  
Jonathan Schmidgall  
Alexis Lopez  
Ian Blood  
Jennifer Wain  
Yeonsuk Cho  
Amy Hutchison  
Hye-Won Lee  
Ahmet Dursun

*December 2018*



# ETS Research Report Series

---

## EIGNOR EXECUTIVE EDITOR

James Carlson  
*Principal Psychometrician*

## ASSOCIATE EDITORS

Beata Beigman Klebanov  
*Senior Research Scientist*

Heather Buzick  
*Senior Research Scientist*

Brent Bridgeman  
*Distinguished Presidential Appointee*

Keelan Evanini  
*Research Director*

Marna Golub-Smith  
*Principal Psychometrician*

Shelby Haberman  
*Consultant*

Anastassia Loukina  
*Research Scientist*

John Mazzeo  
*Distinguished Presidential Appointee*

Donald Powers  
*Principal Research Scientist*

Gautam Puhan  
*Principal Psychometrician*

John Sabatini  
*Managing Principal Research Scientist*

Elizabeth Stone  
*Research Scientist*

Rebecca Zwick  
*Distinguished Presidential Appointee*

## PRODUCTION EDITORS

Kim Fryer  
*Manager, Editing Services*

Ayleen Gontz  
*Senior Editor*

---

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

## RESEARCH REPORT

# Designing a Prototype Tablet-Based Learning-Oriented Assessment for Middle School English Learners: An Evidence-Centered Design Approach

Carol A. Chapelle,<sup>1</sup> Jonathan Schmidgall,<sup>2</sup> Alexis Lopez,<sup>2</sup> Ian Blood,<sup>2</sup> Jennifer Wain,<sup>2</sup> Yeonsuk Cho,<sup>2</sup> Amy Hutchison,<sup>1</sup> Hye-Won Lee,<sup>1</sup> & Ahmet Dursun<sup>1</sup>

<sup>1</sup> Iowa State University, Ames, IA

<sup>2</sup> Educational Testing Service, Princeton, NJ

The tablet project was undertaken in 2013 to explore the creation of an assessment to reveal English language learners' (ELLs') language capabilities in content area classes such as science. Such assessments could ideally be useful in multiple phases of the instructional process by providing information to learners, teachers, and other stakeholders because the technologies allow for efficiencies of task presentation, response gathering, scoring, storing data, and returning results. With the goal of understanding the development costs required to achieve such benefits, the project used an evidence-centered design (ECD) framework to create a prototype assessment for ELL middle school students. In addition to describing the ECD process, the report describes how the project mandate helped to focus the complex processes of ECD and how the activities of ECD came into play in sketching plans for the validity argument.

The ECD process of domain analysis uncovered perspectives on tablet use in middle schools as well as theoretical perspectives for conceptualizing the tablet users' abilities and learning. Findings included a range of uses of mobile technologies in middle school classrooms in 2013, with little uniformity in practices. Based on analysis of common elements in tablet-based communication and learning, a framework for characterizing the relevant strategies and abilities was developed. Findings served in the domain modeling process of defining task design patterns, which in turn provided the basis for planning the conceptual assessment framework (CAF). The CAF also drew upon domain analysis findings, which suggested the concepts required for defining a student model (what is to be inferred about the test taker), an evidence model (the basis for making the inference), and the task model (the features of tasks required for eliciting relevant performance). The assessment implementation process produced a working prototype that served in usability testing and gathering feedback from middle school teachers.

The report describes how ECD guides the design of tablet-based formative assessment tasks that take into account the constructs to be measured and the learning that should result from test taking. The evaluation of the project is undertaken in view of the goal of ECD to design test tasks in a manner that serves in the argument for their use. Success is therefore evaluated in part by an analysis of the quality of content it supplied for an interpretation/use argument for the assessment. Overall, the project highlights include (a) the demonstration of how a mandate delimits the ECD processes; (b) the construct framework encompassing both the linguistic and nonlinguistic resources used to create meaning in a digital environment; (c) the approach used for analysis of a domain with emerging, shifting, and variously used technologies; and (d) the use of an interpretation/use argument in the appraisal of an ECD test design.

**Keywords** Tablet-based assessment; learning-oriented assessment; middle school; English language learners; evidence-centered design; theory of action; validity argument

doi:10.1002/ets2.12232

English language learners (ELLs) in English medium schools face challenges beyond those confronted by students whose native language is English. Therefore, assessments are needed to reveal the language capabilities and needs of ELLs as they pertain to learning in their content area classes such as science. Such assessments are ideally integrated into multiple phases of the instructional process to provide information to learners, teachers, and other stakeholders. The large majority of such assessments will be delivered through technologies to allow for efficiencies of task presentation, response gathering, scoring, storing data, and returning results. Historically, the use of technology has typically been motivated by interest in efficiency. However, over the past 20 years, technologies have gradually become an integral piece of student and teacher communication and learning, both in and out of the classroom. With the changes in practices of language use, the efficiency motive for the use of technology in language assessment is eclipsed to some degree by the need for assessment

*Corresponding author:* C. A. Chapelle, E-mail: carolc@iastate.edu

tasks to be relevant to students' actual language use. The changing role of technology creates a challenging agenda for English language assessment, whose purpose is to deliver efficient and relevant assessments.

This challenge prompted the launch of our Tablet Project in 2013 to undertake research that would lay a foundation for the next generation of English language assessments at the middle school level. The goal was to create a prototype assessment for ELL middle school and, in doing so, to increase understanding of the nature of the challenge itself. A number of broad questions needed to be addressed in order to develop an understanding of the educational uses of mobile technologies that appear to be important for classroom learning, now and in the future.

- How are mobile technologies actually being used in middle school classrooms in the United States (when the research began in 2013)?
- How can the strategies and abilities called for by tablet-based communication and learning be characterized?
- How can particular features of design used in educational apps serve for language learning and assessment?
- How can tablet-based formative assessment tasks be designed to take into account the constructs to be measured and the learning that should result?

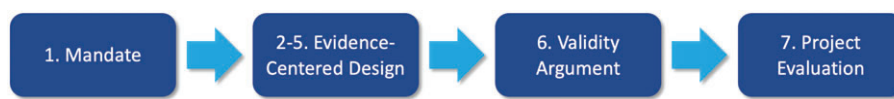
The paths one could take to investigate such issues are numerous, so choices about the direction of the project had to be made in view of the mandate for test design, which was communicated by the initial statement of the purpose of the project expressed by the funder. The mandate should point to certain types of frameworks that need to be identified and developed to carry out the assessment design. The frameworks need to supply useful concepts with an appropriate level of detail to achieve the mandate and to evaluate the success of the project. The position of the mandate in the project is illustrated in the overall schema in Figure 1 showing how we addressed the project.

The project mandate suggested a multifaceted research and development path guided by evidence-centered design (ECD), as described by Mislevy, Steinberg, and Almond (2003) and Mislevy (2011). In this report, the four sections titled Domain Analysis, Domain Modeling, Conceptual Assessment Framework, and Assessment Implementation describe the processes undertaken in ECD, an approach for linking professional knowledge about what is to be tested with the technologies for developing assessments and interpreting the scores they produce. In the Validity Argument section, the connection between the ECD process and the validity argument for the assessment interpretation and use is made. In the Evaluation of the Test Design Process section, we evaluate the success of the project for constructing frameworks, concepts, and prototype materials sufficient to provide a basis for a validity argument that would be needed for such a test.

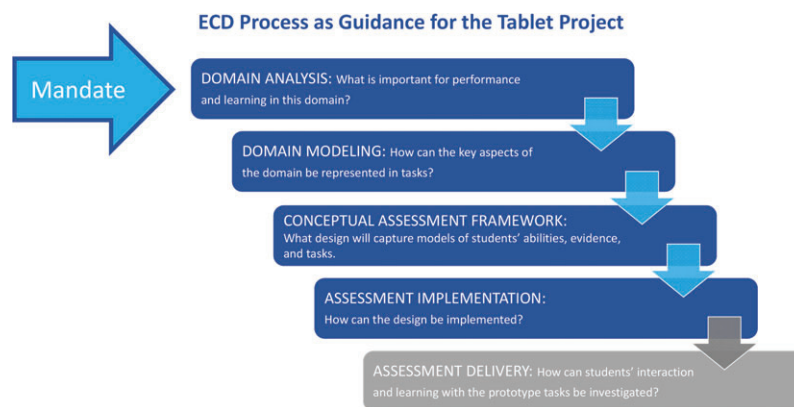
### Mandate-Driven Evidence-Centered Design

ECD is a framework that is intended to build upon traditional judgment-based test design practices undertaken by test developers. The multifaceted professional knowledge that must be taken into account in the design of today's assessments calls for more sophisticated and transparent mechanisms for test design, in large part because of the use of computer technology. More fundamentally, explicit links are needed between design decisions and scores because such links serve as a basis for validation of test score interpretation and use. ECD offers conceptual tools to make explicit the basis for the many test design decisions, but the use of these tools needs to be directed by the mandate for the test design project.

The mandate for a test is defined as the starting point for a test specification, or in our case, the more elaborate test design process. The mandate for a test consists of the "combination of forces which help to decide what will be tested and to shape the actual content of the test" (Davidson & Lynch, 2002, p. 77). The mandate provides guidance about the content that will be discovered and created in the ECD process. The five steps of the ECD process are shown in Figure 2 with the central question that each addresses. Four of the five are introduced briefly here and exemplified by the description of the project in each of the following sections of the report. ECD is conceptualized as five layers leading to delivery, but our project worked through only the first four layers because it was focused on delivering a prototype for research purposes rather than an operational assessment.



**Figure 1** Overall schema showing the sequence for primary drivers of the test design process.



**Figure 2** Schematic diagram of a mandate as antecedent to the layers of the evidence-centered design (ECD) process, consisting of five layers. Adapted from *Evidence-Centered Design for Simulation-Based Assessment* (Figure 1, CRESST Report 800), by R. J. Mislevy, 2011, Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Copyright 2011 by the Regents of the University of California.

In the Tablet Project, the mandate was given by the funder, who wanted to explore the affordances of tablet technologies at the middle school level for formative, or learning-oriented, assessments (LOAs) of academic language performance. In the business of creating and delivering assessments for ELLs, the funder apparently saw potential in the middle school market. Undoubtedly part of a larger initiative, the mandate specified that the assessment had to be delivered on tablets to correspond with the way that tablets would be expected to be used in the classroom for communication and learning in middle schools of the future. The formative classroom assessment would be for providing relevant detailed feedback to teachers and students in addition to total scores, whose meaning would be relevant to classroom instruction. The assessments were also to gather data that would serve in making measurement-based inferences about students' abilities in addition to creating opportunities for students to learn. Because of the exploratory goals of the mandate, decisions about exactly what to assess and what students should learn were not specified. Instead, the detail of the prototype task was to be defined during the course of the project on the basis of the findings of the first stages of the ECD process.

The first step of the ECD process shown in Figure 2, *domain analysis*, had to adopt and develop frameworks relevant to the assessment and learning purposes of the product. In other words, because the mandate was to produce a prototype for an LOA of academic language performance, the frameworks needed to express theoretical perspectives underlying the construct of academic language performance and the hypothesized mechanisms responsible for test takers' learning. The domain analysis also needed to inform the next step, *domain modeling*, which drew upon specifics of the frameworks to describe characteristics of prospective test tasks. The frameworks also provided the terms and concepts needed for the next step in the process, developing the *conceptual assessment framework*, which requires models of students' abilities and specifications of how evidence about their abilities is gathered and summarized, as well as a description of how task characteristics serve in the assessment process. The CAF is rendered as prototype tasks in the *assessment implementation* step. The following sections of the paper summarize how the Tablet Project was developed through each step of the ECD process, which took particular inputs and produced results needed as input for the following step. Because the scope of the mandate included processes only through the development of a prototype, this project concluded with an assessment of the degree to which these four steps yielded sufficient knowledge and data to support aspects of the prospective validity argument.

### Domain Analysis

Domain analysis “concerns gathering substantive information about the domain to be assessed” (Mislevy & Haertel, 2006, p. 7). Exemplifying the type of substantive information investigated for a domain analysis of middle school science, Mislevy and Haertel (2006) further described domain analysis as a process of gathering and analyzing “information about the concepts, terminology, representational forms, and ways of interacting that professionals working in the domain use and that educators have found useful in instruction” (p. 7). Precisely what should be analyzed and how the analysis

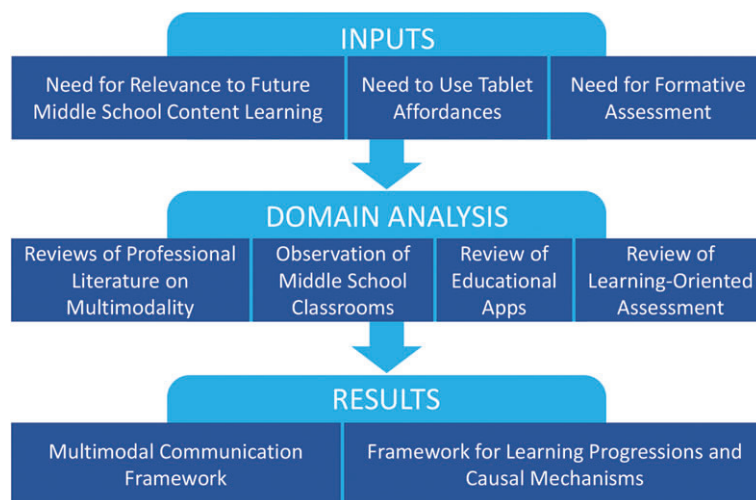


Figure 3 Schematic diagram of inputs, activities, and results for domain analysis.

should be conducted needs to be defined specifically for each test development project, particularly with reference to the test mandate. Three aspects of the mandate served as inputs to the domain analysis process, as illustrated in Figure 3.

### Inputs to Domain Analysis

The mandate indicated that the tablet-based assessments were to be used for formative assessment whose purpose was to provide feedback to teachers and students in addition to total scores whose meaning is relevant to classroom instruction at the middle school level. The three inputs shown in Figure 3 were identified by the mandate: (a) the need for relevance to future middle school content learning, (b) the need to use the affordances of the tablets, and (c) the need for formative assessment that encompasses provision for both measurement and learning. The mandate about relevance to future middle school content learning did not include any direction about the type of content learning or exactly how the assessment was to be integrated into instruction. We therefore had to make these decisions based on the domain analysis. In order to do so in a way that would maximize the relevance of the prototype assessment to middle school classrooms, the domain analysis had to investigate the professional literature and actual classrooms that would help us to understand future technology in learning at the middle school level. Because the assessments were to be delivered as an app on tablets, we needed a way of conceptualizing the mechanisms intended to promote learning through the use of educational apps. For the measurement-based interpretations of performance on the assessment, we needed a means of conceptualizing the abilities engaged by learners as they use tablets for communication and learning in the classroom.

### Four Strands of Domain Analysis

The activities of the domain analysis were carried out through four strands of research that would each provide evidence of what the future might hold for classroom practices and the profession's conceptualization of language, learning, and assessment in such contexts. It was strategically limited in every strand to obtain sufficient information to move to domain modeling despite the very short timeline and limited funding for the project. In keeping with the mandate, we prioritized learning how teachers and researchers conceptualized and talked about learning through technology in content classrooms in U.S. middle schools. The activities of each of the four strands of investigation were described in an interim project report (Chapelle et al., 2015), but they are summarized briefly here along with the results that served as input to the following step in the ECD process.

#### *Review of the Professional Literature on Multimodality*

The review of the professional literature on multimodality was undertaken to discover the concepts and terms needed to describe the type of tasks that learners perform using mobile technologies. The goal was to be able to adopt the terms



and concepts for the construct framework that would describe students' abilities and learning in a way that would resonate with prospective users of the assessment. Teachers could be expected to be educated through the scholarship of the past decades that has developed theoretical perspectives for conceptualizing and studying multimodal communication and learning. In this scholarship, one finds terms such as *new literacies*, which Lankshear, Knobel, and Curran (2012) used to refer to online "fan practices, digital production practices, online reading comprehension, social networking, online communication, video gaming, and [other] new literacies in the classroom" (pp. 3–4). It is recognized that participation in these new literacy events requires a combination of linguistic (e.g., vocabulary and grammar) and nonlinguistic (e.g., mouse clicks, icons, tapping, and swiping) resources to create meaning. For example, what a seventh grader chooses to mean on a given occasion depends in part on the mode he or she chooses to express the meaning and the technology used for the communication. In a classroom activity, students collaboratively nominating what they feel is the theme of a story may express agreement with another student's idea by making an oral linguistic statement, such as "right, I think so, too," a nod of the head, a text message stating, "ok," or a tap on a Like icon.

The expressions *multimodal literacy* and *multimodal communication* are used to refer to the new competencies and practices required for engaging in new literacies that are critical to school learning. Research on multimodality has revealed that not only do communicators use technologies for expressing meaning, but that the technologies actually provide the material resources, which are referred to as *affordances*, for creating meaning. The basic assumption is that meanings are constructed and interpreted through many linguistic and nonlinguistic resources; language is one part of a larger repertoire for conveying meaning (Kress & Van Leeuwen, 2001). Kress (2003) argued that the screen has replaced the book as the central medium of communication. Due to the dominance of the medium of the screen, the modes of image, sound, and color have effects on the form and function of writing. He noted that the many modes of communication that often appear with written text in digital environments (music, color, still and moving images, and speech) all bear meaning and are part of one message. Because of the expanded set of meaning-making potentials and resources afforded by technology, the study of multimodal communication requires an expanded theoretical basis relative to those used to theorize linguistic aspects of communication.

Kress and Van Leeuwen's (2001) theoretical presentation of multimodal communication, which has been central to much of the research on communication and learning, is based on *systemic functional linguistics* (Halliday, 2004; Halliday & Hasan, 1989). Systemic functional linguistics (SFL) is a social-semiotic approach to the study of how people use language to create meaning. SFL provides the analytic categories and concepts to allow researchers to describe the meaning potential within particular contexts as consisting of three layers of metafunctional meaning. The meaning potential includes the linguistic choices (i.e., phonological, lexical, and grammatical) that users can make to express meaning in addition to nonlinguistic resources. Ideational meanings express what is going on through the use of language such as nouns, verbs, and connectors showing logical relations. Ideational functions can also be expressed through nonlinguistic means such as icons, images, physical position, and gestures. Interpersonal meanings are expressed linguistically through, for example, the mood system (e.g., "Would you like to see what I wrote?" expresses a different interpersonal meaning than "See what I wrote?") and a variety of expressions of appraisal. Interpersonal meanings can be expressed digitally by icons indicating likes, or mouse clicks resulting in an email being sent, a blog posted, or a tweet dispatched. Such nonlinguistic moves construe interpersonal meanings through their effect on the audience of the message. In a social-semiotic perspective, the meaning is construed as a result of not only who creates the literacy act but also on the basis of who receives and interprets it. Textual functions are carried out linguistically through choices made about the channel of communication (e.g., written or oral), the language used (e.g., English vs. Chinese), and the positioning of language grammatically and spatially for emphasis, for example. Textual functions are carried out nonlinguistically through the selection of images, sounds, and icons to express meanings and by their placement within a text.

The same basic social-semiotic perspectives toward multimodal literacy form the basis for the Common Core State Standards (CCSS) for schools in the United States. The CCSS in essence acknowledge that new technologies have created new spaces for reading and writing, as well as associated new literacy practices that require new skills, strategies, and dispositions for reading and writing in these spaces (Coiro, Knobel, Lankshear, & Leu, 2008; Leu, Kinzer, Coiro, & Cammack, 2004). From this perspective, then, assessment of new literacy practices requires defining the constructs to be measured in view of an analysis of these skills, strategies, and dispositions. Based on social-semiotic theory, researchers have developed a conception of these new practices and skills, which we drew upon to identify the constructs underlying assessments using tablets, described as one of the results of the domain analysis.

### ***Observation of Middle School Classrooms***

The observation of middle school classrooms was undertaken to develop a practice-oriented perspective on tablet use in middle schools, which served as a reality check on the picture painted in the professional literature. We observed six middle schools in central Iowa identified through the network of school contacts of one of our team members, who obtained invitations from specific schools, teachers, and times of the day when mobile technologies would be used. The classrooms were not intended as a representative sample of typical classrooms in the area. The observations were intended to gather descriptive data as well as to give us a fuller picture of the activities and classroom dynamics taking place when tablets were used in the classrooms. The observations were guided by questions intended to provide a description of the classroom as a whole as well as some specific aspects of tablet use, but they did not attempt to capture the detail of language use because we did not have the resources for recording and analysis of language. Each observation included a brief interview with the teacher to gather data about the class and role of the tablet-based activities.

Overall, the observations confirmed the picture of dynamic classroom interactions and student engagement portrayed in the research literature. The observations also provided a glimpse into the future classroom environment, where the use of tablets and other mobile devices is expected to become part of the daily teaching and learning practices. The classroom observations revealed that with careful planning and clear learning goals, tablets and other mobile devices could be used for teaching in a variety of different subject areas, including English language skills for nonnative speakers, science, and math, as well as literacy and language arts. Observations also revealed that the tablet could be used only for certain purposes and for various periods of times in the class. For example, one sixth-grade science classroom used the Socrative app on iPads and smartphones for a class warm-up activity, and another teacher from the sixth-grade language arts class used the Popplet app on an iPad to expand on the classroom activities they had already started.

Confirming the research reported in journals, the atmosphere in each of the observed classrooms reflected the enthusiasm of teachers and students as they carried out activities using tablets. In contrast to what was reported by others, however, we observed that the use of the mobile technologies in the classrooms was seldom problem free. One common challenge observed across the classrooms was the difficulty the teacher experienced in situating the tasks initially until the students gained autonomy in navigating the apps. The apps appeared to present novel learning environments for learners due to their multimodal design and the range of resources students need to draw upon to accomplish app-based tasks. The challenge of getting students started with novel tasks was not apparent in the research we reviewed, but it was an important factor in the piloting of the tablet assessment that we developed.

### ***Review of Educational Apps***

The review of educational apps provided an up-close view of the features of design that developers exploit to promote learning for this age group for middle school students. The analysis of educational apps complemented the classroom observations, where the tablets remained in the students' hands, leaving us at a distance from the workings of the apps. In order to identify relevant apps for review, we used the Google search engine to conduct a search for online sources dedicated to reviews of iOS- and Android-based apps. We then reviewed online sources (e.g., app review websites, commercial app stores, app review articles) to identify educational apps that were age-appropriate (middle school) and relevant to our target language use domain (language learning or middle school science). The educational apps that we examined had been designed for student learning, and therefore they represented a subset of the types of software described in the professional literature and observed in the classroom, where general purpose apps were also used. We looked at the kinds of learning tasks the apps support, the nature of the linguistic, digital, and semiotic resources students need to control to use the app, what language use contexts the apps are intended for, and the enhancement features the apps have to promote engagement and learning.

The analysis revealed that the apps called upon students' use of linguistic and digital semiotic resources, but with very limited requirements for language production. For example, the linguistic resources required by apps often included discipline-specific vocabulary, whereas digital semiotic resources primarily involved the use of multitouch gestures. Only half the language apps required the learner to produce spoken or written language as part of tasks, and fewer than half the science apps included this requirement. The apps were available across a variety of specific content domains, but the majority were general academic, meaning they could be adapted to learning activities across subject areas. The enhancement features appeared to be critical for task design at the micro level. Table 1 summarizes the enhancement features that



**Table 1** Salient Features of Educational Apps Intended to Promote Learning

Enhancement feature	Definition	Example
Feedback	Information that the task provides about a student's performance or response on a particular task	Verbal, aural, and/or visual information about response correctness
Incentives	Rewards offered for attempting and/or completing tasks, and typically differ based on the quality of performance on the task	Acquiring coins or points; unlocking or accessing new levels or tasks
Control	The degree to which students are enabled to determine how task rewards, goals, and affordances are implemented	
Task difficulty	The manner in which the complexity or difficulty of tasks is manipulated	Difficulty is adapted based on student performance
Environment	The aesthetics, gamelike atmosphere, use of multimedia, animated agents, and interaction between users	A gamelike or aesthetically pleasing environment; multimedia such as video; a narrative, immersive, and authentic environment; animated agents or avatars in the environment

were used to promote learning. We were particularly interested in these features in view of the previous research that has shown that their implementation may lead to increased student engagement, motivation, self-regulation, or self-efficacy (McNamara, Jackson, & Graesser, 2010). Given the goal of creating a classroom-based formative assessment that engages learners in a tablet environment, we incorporated many of these features into our task design.

### **Review of Learning-Oriented Assessment**

The review of LOA was undertaken to conceptualize an approach to the formative assessment aspect of the mandate. From this literature, we identified four qualities that an LOA should display to serve as a tool for promoting the acquisition of targeted knowledge or skills, and perhaps learning strategies. First, LOA tasks should stimulate sound learning practices by encompassing worthwhile educational value in and of themselves (Bennett, 2011; Carless, 2007). This principle requires assessment tasks to “embody the desired learning outcomes” (Carless, p. 59), which includes promoting desired “learning dispositions” (presumably, relevant metacognitive strategies and skills) and using practical, real-world situations.

Second, LOA tasks should be designed to enhance learners' motivation and promote student engagement by stimulating their interest. Recent work on motivation includes the role of the instructional context and importance of specific features of learning tasks that can promote motivation. Similarly, *interest* refers to the extent to which underlying needs or desires of the learner are activated. It may be further differentiated according to *individual* or *situational interest*; the former implies a personal investment in a topic or domain, whereas the latter may arise based on an interaction among contextual features. Thus, even when the topic or content domain may not activate individual interest, contextual features of the task (e.g., gamelike elements) may drive situational interest.

Third, LOA tasks should provide descriptive and timely feedback. Provision of feedback to learners is among the most important conditions for second language acquisition (SLA) that is recognized in SLA research (Li, 2010; Lyster & Saito, 2010; Valezy Russell & Spada, 2006). Feedback is theorized as critical to SLA because it provides learners with the opportunity to contrast their own knowledge with the actual learning targets. In SLA, typologies of feedback typically distinguish between prompts and reformulations, depending on whether feedback provides a corrected form of the learner's utterance (Ranta, Lyster, & DeKeyser, 2007). Feedback on linguistic performance can take many different forms, but all share the characteristic that they can be defined as information about the students' current learning or performance in relation to a specific goal or standard (Nicol & Macfarlane-Dick, 2006). In other words, feedback is information about where the students are in their efforts to reach a goal (Wiggins, 2012). McNamara et al. (2010) argued that feedback is vital to the learning process and that feedback is a critical aspect of learning environments.

Fourth, LOA tasks should create opportunities for students to reflect on their performance goals and to engage in self-assessment by applying evaluative criteria to their performance. This quality is based on the finding that students who set appropriate learning goals employ strategies to achieve them, and those who reflect on learning outcomes tend to perform better and feel better about their learning. These behaviors are associated with self-regulated learning (SRL),

1	construct meaning from oral presentations and literary and informational text through grade-appropriate listening, reading, and viewing	Standards 1 through 7 involve the language necessary for ELLs to engage in the central content-specific practices associated with ELA & Literacy, mathematics, and science. They begin with a focus on extraction of meaning and then progress to engagement in these practices.
2	participate in grade-appropriate oral and written exchanges of information, ideas, and analyses, responding to peer, audience, or reader comments and questions	
3	speak and write about grade-appropriate complex literary and informational texts and topics	
4	construct grade-appropriate oral and written claims and support them with reasoning and evidence	
5	conduct research and evaluate and communicate findings to answer questions or solve problems	
6	analyze and critique the arguments of others orally and in writing	
7	adapt language choices to purpose, task, and audience when speaking and writing	
8	determine the meaning of words and phrases on oral presentations and literary and informational texts	Standards 8 through 10 hone in on some of the more micro-level linguistic features that are undoubtedly important to focus on, but only in the service of the other seven standards.
9	create clear and coherent grade-appropriate speech and text	
10	make accurate use of standard English to communicate in grade-appropriate speech and writing	

**Figure 4** English Language Proficiency (ELP) Standards and their organization as presented in the ELP Standards document. From *English Language Proficiency (ELP) Standards* (p. 4), by the Council of Chief State School Officers, 2014, retrieved from [https://ccsso.org/sites/default/files/2017-11/Final%204\\_30%20ELPA21%20Standards%281%29.pdf](https://ccsso.org/sites/default/files/2017-11/Final%204_30%20ELPA21%20Standards%281%29.pdf)

defined as the “self-directive process by which learners transform their mental abilities into academic skills” (Zimmerman, 2002, p. 65). Adaptive self-regulatory processes are related to better academic performance, engagement in learning, positive self-beliefs, and higher motivation (e.g., Pintrich, 2000; Pintrich & De Groot, 1990; Pintrich, Roeser, & De Groot, 1994; Schunk, 2005; Wolters, Yu, & Pintrich, 1996), which has led some educational psychologists to argue that fostering self-regulated learning should be a major focus of education (Boekaerts, 1997). Our LOA should be evaluated based on evidence about the degree to which students display increases in SRL as they progress through the assessment tasks.

Overall, these LOA features need to be crafted for any specific assessment according to a theory of action that specifies the intended relationships among the design elements of the assessment tasks, the intended learning behavior, and desired outcomes. Such a theoretical framework is needed as guide for assessment developers to integrate the measurement and learning aspects of the assessment at the planning stage. A learning-oriented assessment framework in this context would also require a well-founded basis for characterizing the learning progressions that would be meaningful to the students and teachers in middle school classrooms. Learning progressions are statements about the specific abilities that must be mastered for students to ultimately be able to achieve the complex learning outcomes that are the pillars of the curriculum. As Popham (2008) put it, “Learning progressions, in an almost literal sense, become the maps that provide guidance on how best to carry out formative assessment” (p. 29). Linguistic abilities are prime examples of the building blocks that students must master.

Linguistic learning progressions have been developed to guide instruction for ELLs who are studying in a curriculum based on the CCSS. Each of the English Language Proficiency (ELP) Standards (Council of Chief State School Officers, 2014) has a corresponding set of progressions that describe what students should be able to do as they build their language skills toward each of the standards. The standards themselves are focused on helping students develop contextualized language proficiency for communicating in ways called for in school settings and for achieving academic success within specific disciplines. Figure 4, from the ELP Standards document, lists the 10 standards and notes how they are further organized to represent each standard’s importance to ELLs’ participation in the practices called for in content classes based on CCSS.

The ELP Standards specify that students must not only have strong receptive language skills, but must also be able to produce oral and written language for varying tasks, audiences, and purposes. In regard to their receptive language skills, students need to understand the meaning conveyed in oral face-to-face presentations as well as from multimedia presentations such as videos, podcasts, and slideshow presentations. Students must use their productive language skills to be able to interact with each other through oral and written language. For example, ELP Standard 2 (as shown in

Figure 4) states that students should “participate in grade-appropriate oral and written exchanges of information, ideas, and analyses, responding to peer, audience, or reader comments and questions.” In a digital society, both production and interpretation of language necessarily involve the use of digital technology. In a technology-rich classroom, language use could include composing blog posts, podcasts, videos, and websites. It could also mean interacting through social media such as Twitter or responding to a peer’s digital composition. Whatever the task, the use of digital technology often requires that students combine the receptive, productive, and interactive modalities of language in a variety of ways to create meaning, as described in the research on multimodal communication. The way meaning is made through digital technology must be taken into account in developing tasks for learning and assessment of the language standards.

## Results From the Domain Analysis

The intended outcome of domain analysis is to lead test designers “to understand the knowledge that people use in the domain, the representational forms, characteristics of good work, and features of situations that evoke the use of valued knowledge, procedures, and strategies” (Mislevy & Haertel, 2006, p. 7). These understandings are not intended to translate directly into assessment designs, but rather to “presage the entities and structures that appear in subsequent layers” (p. 7). In the Tablet Project, the results from the domain analysis included our own increased understanding of how tablets are used for learning through both general purpose software and with apps designed specifically for learning. We developed an understanding of the way that multimodal communication is viewed in applied linguistics and education. A number of results were presented from each of the lines of inquiry throughout this section; the following builds on those with the two frameworks developed on the basis of the results. We developed a multimodal communication framework and a theory of action framework for learning to characterize the abilities required of students performing tablet-mediated tasks and the mechanisms underlying learning during task performance, respectively. The two frameworks express the representational forms and features of situations that are used for subsequent steps in task design.

### **Multimodal Communication Framework**

The review of the professional literature on multimodal communication in middle school content learning helped to develop a construct framework for multimodal communication. This framework is not intended to define the construct of the assessment to be developed, but rather to provide a basis for defining specific test constructs that are intended to be relevant to abilities in tablet-mediated language use in middle school content classes. Accordingly, the construct framework presents the range of contexts, tasks, semiotic functions, and semiotic resources that are expected to come into play in the classrooms of the future. Any particular assessment needs to be specified in more detail depending on its intended purpose. Depending on the purpose, test tasks requiring abilities to perform in these contexts and tasks may be considered to assess construct-relevant abilities.

Past work on multimodality has not been connected with language assessment, but the study of multimodality does have a basis in *functional linguistics* (the study of language from the perspective of how it conveys meaning in context). Functional linguistics, to some degree, has also influenced communicative language frameworks used in assessment. In particular, the draft ELP framework document to be used by ETS to guide all ELP test development for schools identifies functional linguistics as a basis for the conceptualization of communicative language proficiency. Moreover, the ELP Standards for ELLs that will be important for ultimate score interpretation are also functionally based. The *TOEFL*<sup>®</sup> program at ETS has developed a framework for a system of assessments for school-based language. The framework document, *Building a Framework for a Next-Generation English Language Proficiency Assessment System* (Wolf et al., 2014), adopts theoretical frameworks from applied linguistics to describe the language abilities. The authors introduce the basis for their framework: “We propose that communicative competence models in second language acquisition and learning, academic English language literature, and various standards (both ELP and academic content standards) should be reviewed to define the ELP construct” (Wolf et al., 2014, p. 3). The broad functional orientation of the ELP framework is evident in its conceptualization of “two broad purposes for communicative language use in K–12 school settings.” One is “accessing academic learning in school contexts using foundational and higher order language skills,” and the other is “engaging with peers, teachers, and staff in school contexts that are not strictly content learning-focused, using foundational and higher order language skills” (Wolf et al., 2014, p. 6). The framework concentrates on the use of language for accessing content learning and engaging with peers.

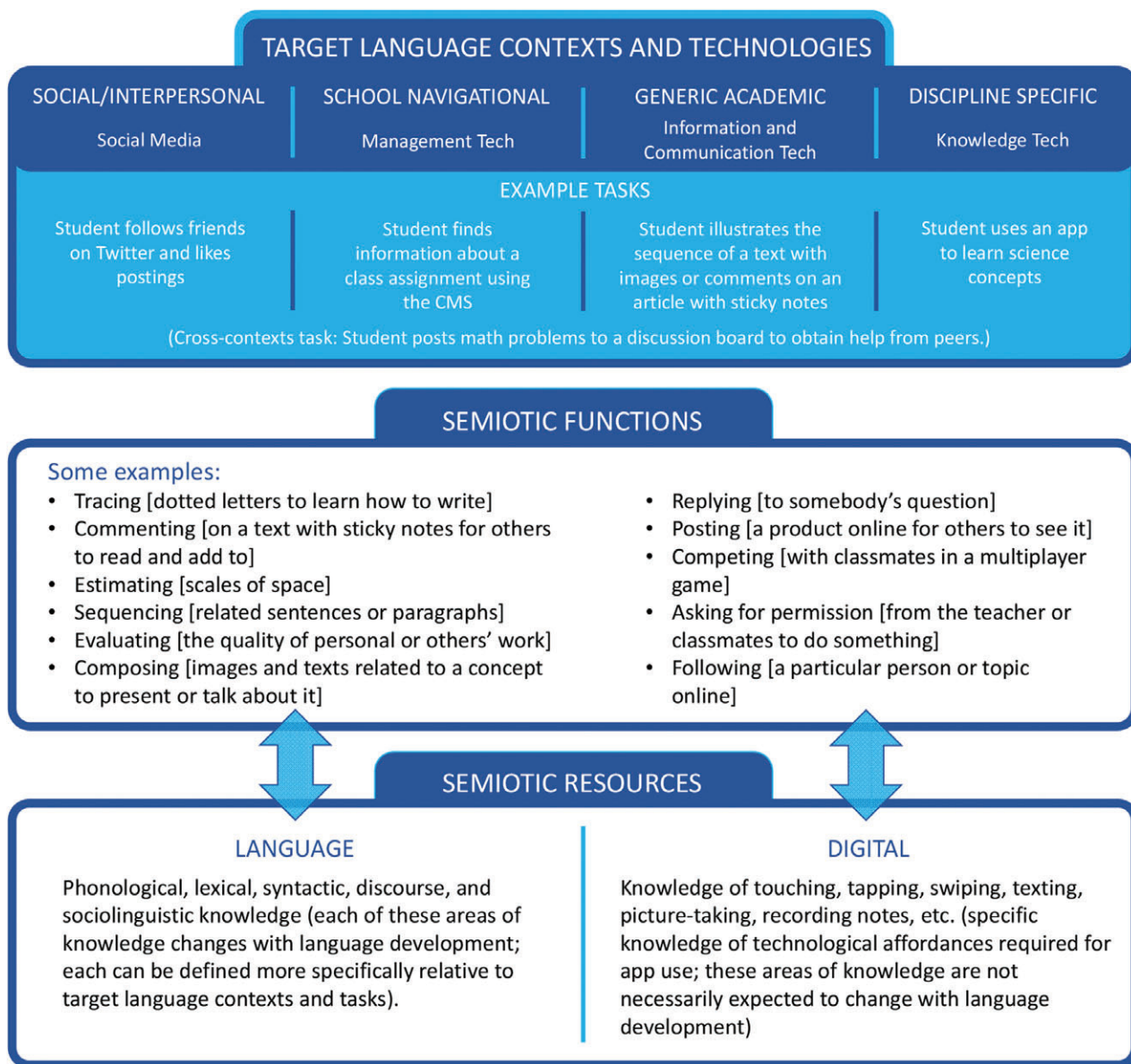


Figure 5 Framework for a multimodal communication construct. CMS = content management system.

This functional orientation is useful, but to capture the critical features of communication and learning through the use of tablets, it needs to include the types of technologies that mediate language use in addition to the nonlinguistic resources such as images, graphics, sounds, and various types of movement of the hand and device. The framework for a multimodal communication construct shown in Figure 5 is intended to help test developers conceptualize the constructs of multimodal communication and learning as they are called on in middle school classrooms using tablets. The framework was designed to be consistent with the ELP construct framework presented in the framework document for the Next-Generation English Language Proficiency Assessment System. It is not intended to specify a construct to be measured by a specific test, but rather as a framework identifying aspects of a construct of multimodal communication that test developers may want to consider in specifying the construct of interest for any particular assessment.

The Target Language Contexts and Technologies heading at the top of Figure 5 refers to the types of situations in which students in middle school would be expected to make meaning and the technologies that may be selected to mediate the communication and learning. Four types of situations are defined as social-interpersonal (e.g., students communicating with one another), school navigation (students engaging with the rules, requirements, and procedures of



school-based work), generic academic (students participating in genres of communication that cut across specific areas, such as expressing opinions about a reading in writing), and discipline specific (students using learned disciplinary knowledge to complete a task). Situations are important because the situation influences the types of tasks that may occur, the functional meanings that the student would need to create, and therefore, the resources required to create the meaning. For each type of situation, certain classes of technologies are likely to play a role, as illustrated in the example tasks directly under the respective situations. Therefore, students need to be able to use the corresponding digital resources in order to participate. These contexts can be used as a macro organization for the tablet tasks. Some tasks will be easy to classify as fitting within a particular situation, but others will have multiple purposes and therefore will not fit neatly into one category. The Example Tasks section in Figure 5 illustrates tasks that students accomplish through the use of apps in each of the contexts.

*Semiotic functions* are the functional meanings that the student needs to be able to carry out. Language is one way of making meaning, but other ways include the use of images and drawings or using signals within an app to indicate functional meaning, such as a request (e.g., tapping on a word to get a definition or to hear it pronounced). These functions can overlap to some degree with language functions, but there are some that are unique to app use as well.

*Semiotic resources* refer to the knowledge required to perform functions. In Figure 5, the interdependence of semiotic resources and functions is indicated by the double-headed arrows between these components. Semiotic resources can be thought of as tools that people use to create meaning. They consist of both language (e.g., words and grammar) and nonlanguage (e.g., gestures, facial expressions, images, and screen tapping). These linguistic and digital resources are intermingled during task completion in a manner that makes it impossible to separate them out in real performance, but they are separated for analysis here. In our framework, we include the nonlanguage resources that people use to make meaning in the digital environments created by tasks that use apps (e.g., tapping, texting, and taking pictures).

### **Framework for Learning in a Learning-Oriented Assessment**

The review of LOA resulted in a theory of action that specifies how the assessment design is supposed to promote students' learning within the larger system of the classroom. The framework shown in Figure 6 maps out how the tablet task design features (LOA components) are supposed to affect students' learning as they work on the tablet tasks. It states that the task components—including diagnostic feedback, rewards for task completion, gamelike environment, immersive environment simulated with a narrative, and avatars and animated agents—should engage certain hypothesized action mechanisms. The action mechanisms are the actions and processes that take place during optimal use of the tablet task, including teachers targeting students' needs on the basis of their use of the feedback, learners' focus on appropriate goals, and learners' use of the assessment. Each of these actions is expected to have certain desirable intermediate effects, including improvements in teaching, learner self-efficacy, self-regulation, and engagement. These intermediate effects, in turn, are intended to have ultimate positive effects on teaching and learning. Specifying the action mechanisms and their intended intermediate and long-term effects provides a basis for developing relevant microlevel evaluation procedures for the assessment.

Overall, the activities of the domain analysis yielded two frameworks offering a systematic conceptualization of the findings to be used in the next steps of the domain analysis. In particular, the multimodal communication framework is useful for specifying the student model required for the CAF. The CAF also requires specification of task features that are included in the theory of action framework. Both frameworks also contribute to the specification of aspects of the validity argument and therefore play a critical role in the evaluation of the Tablet Project. These and other uses of the frameworks will be noted in the following parts of the paper, but in the next section, they are put to work for the next step of the ECD process, domain modeling.

### **Domain Modeling**

Domain modeling is the second phase of test development in the ECD process. Its purpose is to “lay out what an assessment is meant to measure and how and why it will do so” (Mislevy & Haertel, 2006, p. 8). Central for the domain modeling process is the development of structures that document the links between the rationales for test design and the concrete plans that specify the parameters for the test tasks. Mislevy and Haertel called these rationale-based concrete plans “design patterns, which help assessment designers think through substantive aspects of their assessment argument, in a structure

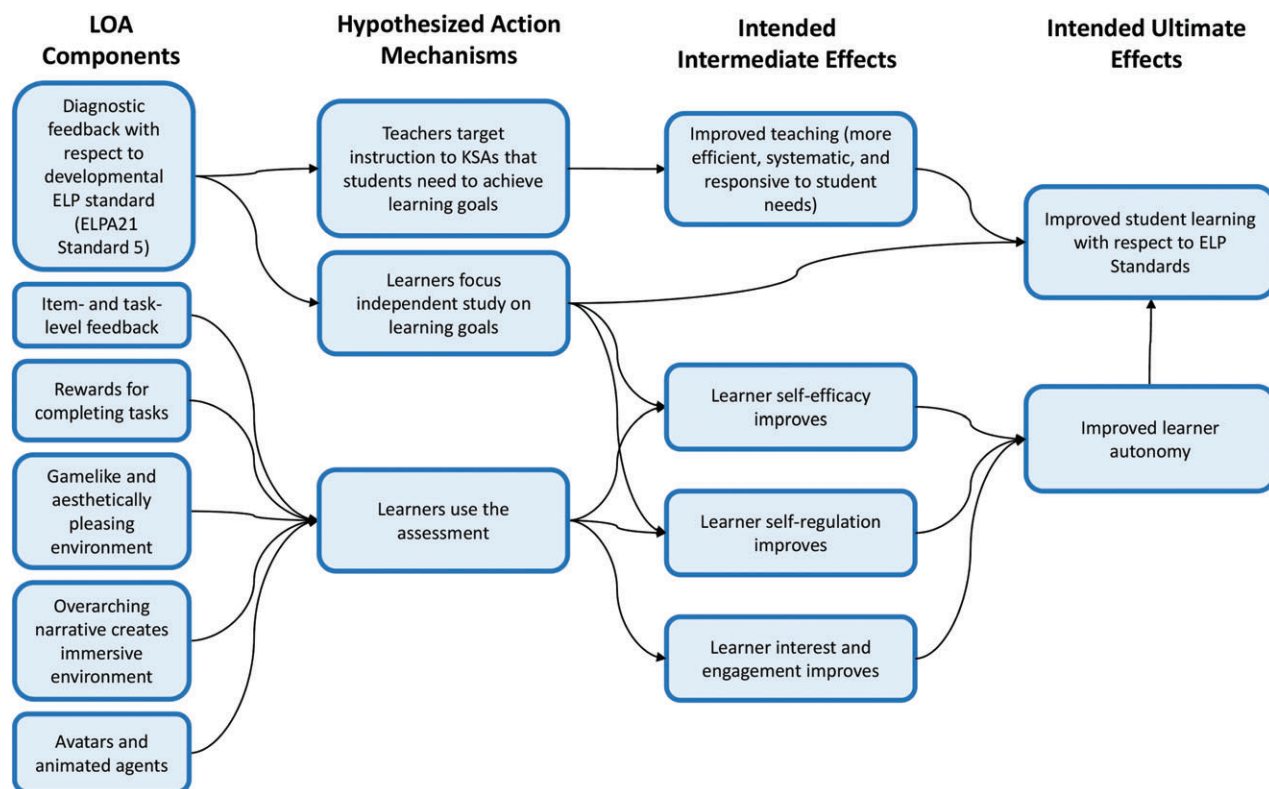


Figure 6 Theory of action for LOA. Adapted from *Research Rationale for the Keeping Learning on Track® Program*, by Educational Testing Service, 2009, Princeton, NJ; Author. Copyright 2009 by Educational Testing Service. KSA = knowledge, skills, and abilities; LOA = learning-oriented assessment.

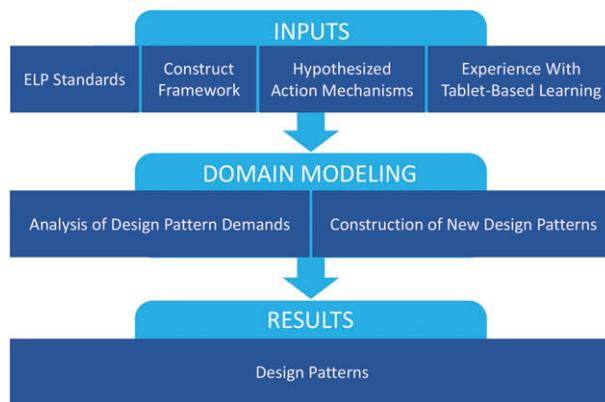


Figure 7 Schematic diagram of inputs to domain modeling and its results.

that is useful across specific domains, educational levels, and assessment purposes, and lead to the more technical work in the next layer” (p. 8). In domain modeling, the rationales need to come from the substantive knowledge of the domain that is developed in domain analysis. Therefore, the results of the domain analysis serve as inputs to the domain modeling process, as illustrated in Figure 7. The links are made between these and the design patterns through the process of domain modeling, which consists of analyzing the demands of the design patterns and constructing new design patterns based on the domain analysis. The results from this process are the design patterns containing sufficient specification to serve in the next step of the ECD process.



## Inputs to Domain Modeling

Inputs for the domain modeling process need to be interpreted as rationales that provide the reasons for certain decisions about the design patterns. Four aspects from the domain analysis served as rationales for the design patterns described in this section. The ELP Standards provided the rationale for the intended interpretation for the score(s) from each of the assessments. As a communication tool, the ELP Standards are intended to allow teachers across various contexts to conceptualize the dimensions of linguistic ability that their students should be developing. The interpretation of test scores needed to fit into this scheme in order to provide scores that would help teachers understand how students were doing with respect to those targeted abilities. Each test then had to yield a score or scores that teachers would see as useful for informing their future instruction. This framework provided a means for narrowing down the scope of the prototype project to address a single standard for a particular level, which we chose to be ELP Standard 5 (see Figure 4).

The construct framework provides the rationale for conceptualizing the range of potential learner and task factors that are important for assessing the targeted abilities. The construct framework complements the ELP Standards by providing a more detailed conception of what is to be measured, which is needed for creating design patterns because the decisions about specific features of tasks require a detailed understanding of what is measured in addition to the ELP Standards statement, which is intended for communication about learning. The hypothesized action mechanisms provide the rationale for inclusion of particular features of the tasks in the design patterns by stating hypotheses about how the features are supposed to prompt optimal performance from test takers as well as promote learning. Such features include feedback, which is intended to engage test takers in the tasks and increase their understanding of their own abilities. Finally, experience with tablet-based learning provided the rationale for specifying the particular affordances and general learning scenarios that were plausible for the assessment in view of those used in classroom learning. The specific uses of these inputs as rationales for design patterns were made clear through the first part of the domain analysis, which consisted of analyzing the demands of the design patterns.

## Activities of Domain Modeling

The precise activities of the design modeling process can be carried out in a variety of ways, depending on the extent and findings of the domain analysis. Our domain analysis had provided a wealth of ideas and two complex frameworks that we wanted to bring to bear on the process of creating the design pattern that would be the outcome of this step. We therefore began with an analysis of what was needed to create the design patterns by analyzing examples of design patterns prior to constructing our own.

## *Analysis of Design Pattern Demands*

The design patterns as specified by Mislevy and his colleagues are defined and exemplified in several test design documents for other assessments. Mislevy and Haertel (2006) defined design patterns as consisting of the attributes shown in Table 2. We examined these components of the prototypical design pattern to discover how we could use the results of the domain analysis to yield the specifications needed for each of the seven attributes. Our design patterns include these elements, with some modifications and additions that were needed to communicate the particulars of the design patterns for the Tablet-Based Test of Academic Language Performance. The left two columns of Table 2, adapted from Mislevy and Haertel (2006), provide a definition of each of the attributes included in the design patterns. The four columns on the right indicate the inputs to the domain modeling process that served as a rationale for each attribute in our design patterns.

The implications of these rationales are evident in the design patterns that we developed. The first five all draw upon what the intended interpretation of the test score(s) should be, or what the test is intended to measure. First, the titles are consistent with the wording in the ELP Standards, described above. Second, the summary needs to express the intent of the assessment in a way that communicates to teachers through the terms used in the ELP Standards, which indicate the performance that the test taker is supposed to display to achieve the standard. Third, the rationale in the design pattern requires a statement of the relevant context in which the score is intended to have meaning. In a general sense, this is stated in the mandate, but more specifically, the construct framework includes a more detailed specification of the context in which the scores are relevant. Fourth and fifth, the construct framework provides the basis for an analysis of the knowledge, skills, and abilities (KSAs) required for performance in the contexts specified, as well as additional KSAs. In view of the

**Table 2** Design Pattern Attributes and Definitions

Attribute	Definition/purpose	ELP Standards	Construct framework	Theory of action	Experience with tablet-based learning
1. Title	Brief preview of the abilities that the design pattern is to address expressed in terms that are relevant to the prospective test users.	X			
2. Summary	Summarizes the intent of the tasks defined in the design pattern.	X			
3. Rationale	The connection between the focal abilities and what people do in what kinds of circumstances beyond the testing context.		X		X
4. Focal KSAs	The primary KSAs targeted by this design pattern.		X		
5. Additional KSAs	Other KSAs that may be required by tasks written under this design pattern even though they are not the intended targets of measurement.		X		
6. Potential observations for evaluation (work products)	Some possible things one could observe students saying, doing, or creating that would provide evidence about the KSAs.		X		X
7. Potential observations for evaluation	Features of the work products that constitute the evidence.		X		
8. Characteristic features of task	Aspects of assessment tasks that are necessary in some form to evoke the desired evidence of ability and promote learning.		X	X	
9. Features of tasks that vary across levels	Aspects of assessment situations that can be varied in order to shift difficulty or focus, while maintaining what is needed to evoke the targeted evidence.		X		
10. Features of tasks specific to certain levels	Aspects of assessment situations that must be varied to elicit observations while maintaining what is needed to evoke the targeted evidence.		X		X

*Note.* Portions of this table are adapted from “Implications of Evidence-Centered Design for Educational Testing,” by R. J. Mislevy & G. D. Haertel, 2006, *Educational Measurement: Issues and Practice*, 25(4), Table 2. Copyright 2006 by National Council on Measurement in Education. KSAs = knowledge, skills, and abilities.

focus on the language resources specified in the construct framework, the additional abilities for many tablet-based tasks include the nonlinguistic semiotic resources required for operation of the tablet.

The following two attributes (6 and 7) are about the types of observations that can be made. The potential work products, or performances, need to be specified in view of the analysis of the affordances of the technology in implementing the desired tasks. Therefore, such specifications were developed from our experience with tablet-based learning. These come from an analysis of the work products in view of the construct framework and the specific ELP Standard targeted. The potential observations to be made from the work products come from the specifics of the intended construct meaning. The construct framework provides guidance by including the areas of language competence that may be relevant to the intended construct meaning. What is done in creating the task patterns is to identify aspects of performance that are expected to appear in the work product and that can be used to infer aspects of the construct.

The final three attributes of the task patterns (8–10) are about the tasks themselves. The characteristic features of the task were defined based on our analysis of the tasks that would be needed to yield the features of performance from which the abilities of interest could be inferred. We used the construct framework, hypothesized action mechanisms,

and experience with tablet-based learning to develop the characteristic features. To identify features that should be varied systematically in tasks eliciting observations at different levels for each standard, we used the ELP Standards, the construct framework, and our experience with beginning to design the prototype task. The features of tasks specific to certain levels come from the ELP Standards and the construct framework. This analysis of the requirements for the task patterns allowed us to make connections between the results of the domain analysis and our task patterns with the intent of developing a defensible basis for the measurement model.

### ***Construction of New Design Patterns***

Construction of new design patterns to capture the critical task features was undertaken as an interactive process that drew upon both the results of the domain analysis as described above and our simultaneous experience with task design. The nonlinearity of the ECD process became evident at this stage, when it would have been impossible to write the specifications for design patterns without also having some concrete experience with design of a specific task, even though such development is described in ECD as occurring later in the process.

The first step was to use the ELP Standards as the basis for identifying a learning outcome for which our prototype assessment could serve to provide scores indicative of students' ability level in middle school. There was no mandate-related reason for choosing one learning outcome over another. We chose to develop the prototype for ELP Standard 6–8.5, which is stated as follows: “The student will conduct research and evaluate and communicate findings to answer questions or solve problems.” Identification of this targeted outcome provided sufficient information to begin to define elements of the design pattern, including the title and summary.

The first two attributes, title and summary, come directly from the outcomes that are stated in the ELP Standards. However, each standard does not necessarily map directly to one task pattern. Analysis is needed to determine how many task patterns should be developed to assess and teach material needed to achieve each standard. For example, ELP Standard 6–8.5 included both conducting research and communicating findings. We determined that these two aspects of the standard each warranted its own assessment. This decision was made on the basis of our attempts to develop assessment tasks that included both. This experience in prototype development was needed to understand what was involved in each part of the standard.

The rationale needs to include the context of language use, and therefore we drew upon both the construct framework, which specified potential domains of academic language use, and our experience with tablet-based learning in middle school. The domain analysis results showed that language ability is called upon for performance across a range of school situations. The specific types of language performance required depend in part on the context of language use. A lot of information is accessible through print and digital sources, but an individual must have the ability to access the information and assess relevance and credibility of each source. Using available sources, students collect information useful for answering questions they have or solving problems given to them. They also communicate findings to one another.

The focal KSAs and the additional KSAs were identified primarily through the use of the construct framework. Within the construct framework, the focal KSAs fell within the general academic target language context. In that context, the phonological, lexical, syntactic, discourse, and sociolinguistic knowledge required for performance on tasks that require searching and communicating was of interest. The additional KSAs are any of the linguistic resources required for social-interpersonal interaction, school navigational purposes, or discipline-specific tasks. Additional knowledge would also include the digital resources identified in the construct framework, which may be required for task performance but are not part of the intended interpretation of the test scores. The semiotic functions are included in the focal KSAs only insofar as they are accomplished through language resources and are the types that would be used on generic academic tasks.

To identify potential work products, we analyzed the nature of the data needed to draw inferences from performance to students' ability to gather and communicate information in general academic tasks. Performance had to be based on work products coming from multiple engagements with a coherent task in order to make inferences about gathering and communicating in a general academic context. Even though task-based learning was not investigated as part of the domain analysis, models and ideas for such tasks came from our experience in working with task-based learning, where students work with scenarios involving problem solving and communication. Potential observations for evaluation needed to be identified on the basis of the intended construct to be measured. The construct framework was used to identify the aspects of performance that would be relevant for observation.

**Table 3** Enhancement Features Theorized to Promote Engagement and Motivation, Which in Turn Promote Learning

Enhancement features	Definitions	Examples	Examples of implementation in the prototype
Item- and task-level feedback	Information that the task provides about a student's performance or response on a particular task	Verbal, aural, and/or visual information	Instantaneous: points and stars (visual), dings and buzzes (aural) Delayed: medals to summarize task performance (gold, silver, bronze); badges to summarize development progress
Rewards for completing tasks	Rewards offered for attempting and/or completing tasks, typically differ based on the quality of performance on the task	Acquiring coins or points; unlocking or accessing new levels or tasks	Acquisition of points; acquisition of medals; aesthetic incentives such as uncovering a mural in Diagnostic Task 1 or changing the color of the river in Diagnostic Task 2
Environment	The aesthetics, gamelike atmosphere, use of multimedia, and animated agents	A gamelike or aesthetically pleasing environment; multimedia such as video; a narrative, immersive, and/or authentic environment; animated agents or avatars in the environment	Comic book, film noir art aesthetic; use of video and audio in sources; use of an overall narrative; Citytown's attempt to win the Best Town Contest); customizable detective avatars

The characteristic features of tasks were designed on the basis of the intended construct to be measured in addition to the hypothesized action mechanisms that were intended to promote learning. In particular, the tablet task design features (referred to as LOA components in Figure 6) are affordances of the tablet technology intended to promote student engagement with the assessment tasks. These are referred to as enhancement features in Table 3, which includes examples of each type. The features of tasks that vary across levels and features of tasks specific to certain levels were identified on the basis of the learning progressions in the ELP Standards. For ELP Standard 6–8.5, our analysis determined that this standard required two sets of tasks types, one requiring students to gather information (i.e., research) and a second requiring them to communicate findings.

### Results of Domain Modeling

The results of the domain modeling phase were the draft design patterns for each of the ELP Standards (Chapelle *et al.*, 2015). Because each standard contains multiple constructs that can be assessed, more than one design pattern was developed for each of the standards. The initial sketches for all the standards provided a starting point for future assessment development. However, until task development is undertaken, they remain in draft form. In contrast, the design patterns for the prototype assessment were developed more completely through a process of using results of the domain analysis while exploring task development. Tables 4 and 5 present the results of that process, which include two design patterns that specify the prototype assessment task for ELP Standard 5 for Grades 6–8 at the beginning level. Each design pattern contains attributes that describe the important characteristics of the tasks. For each attribute, the specific design pattern is defined by its values. For each of the values, additional notes are provided in the Comments column. Comments typically include the source of the specific value, but can also contain other notes that provide additional information to be used in task design.

In ECD, the design patterns are intended to link the rationales developed through domain analysis to the design of the assessments. The design patterns and their process of development described in this section accomplish this goal. We also noted that creating the design patterns on the basis of the domain analysis alone was not possible. The results for the domain analysis were rich in encompassing many ideas about what assessment tasks could be, but the domain analysis process, even in view of the mandate, did not provide sufficient constraints to allow a principled description of task patterns. Instead of moving sequentially from domain analysis to domain modeling, it was necessary to gain some experience in sketching materials for potential tasks. In order to identify some concrete specifications for the design

**Table 4** “Gathering Information” Design Pattern

Attribute	Values	Comments
Title	Gathering information	Title comes from the ELP 6–8.5 Standard. “Gathering information” is the Level 1 function that forms the basis for the more complex ELP Standard 6–8.5. ELP Standard 6–8.5 is “conducts research and evaluates and communicates findings to answer questions or solve problems.”
Summary	This design pattern concerns gathering information from sources to answer questions or solve problems. It is the most basic function leading up to ELP Standard 6–8.5.	Our review of the literature and observation of middle school classes revealed the role of language abilities in using technologies for conducting research. The construct framework includes academic contexts as necessary for specifying construct meaning.
Rationale	Language ability is called upon for performance across a range of school situations. The specific types of language performance required depend in part on the context of language use. Information is accessible through print and digital sources, but an individual must have the ability to access the information and assess relevance and credibility of each source. Using available sources, students collect information useful for answering questions they have or solving problems given to them. Gathering information from given or found sources to answer questions or solve problems on the academic topic chosen for the assessment	The construct framework theorizes functions as the primary category for conceptualizing knowledge skills and abilities. Particular functions are called upon by certain tasks and in turn require specific semiotic resources (including both linguistic and nonlinguistic). The construct framework theorizes knowledge skills and abilities as working together to accomplish the functions called upon by tasks in particular contexts. Test takers’ task familiarity and knowledge inevitably come into play in successful task completion.
Focal KSAs	<ul style="list-style-type: none"> <li>● Familiarity with task</li> <li>● Subject-area knowledge</li> <li>● Nonlinguistic semiotic resources for interacting with the technology</li> </ul>	Review of research, class observations, and analysis of educational apps revealed typical and possible work products.
Additional KSAs	<ul style="list-style-type: none"> <li>● Identification of questions to be answered / problems to be solved (e.g., describing the problem, correctly identifying the key among distractors)</li> <li>● Successful location of relevant information (e.g., identifying information as relevant or irrelevant)</li> </ul>	<ul style="list-style-type: none"> <li>● Identification of questions to be answered / problems to be solved (e.g., describing the problem, correctly identifying key among distractors).</li> <li>● Successful location of relevant information (e.g., identifying information as relevant or irrelevant).</li> </ul>
Potential observations for evaluation (work products)	Results of searching, identifying, or locating information, which can be marked or moved into a designated area.	
Potential observations for evaluation (implications for potential rubrics)	<ul style="list-style-type: none"> <li>● Identification of questions to be answered / problems to be solved (e.g., describing the problem, correctly identifying the key among distractors)</li> <li>● Successful location of relevant information (e.g., identifying information as relevant or irrelevant)</li> </ul>	The task should be presented to appear like the tablet-based tasks that the students perform in their science classes. The software should include gamelike features to help maintain interest and create incentives. Feedback must be provided to promote learning.
Characteristic features of tasks	Academic situations where answering questions or solving problems is required, and a set of sources to refer to in order to answer the identified questions or solve the identified problems. Tasks require students to use strategies for searching for new information to answer the identified questions or solve the identified problems. Feedback must be provided.	ELP Standards further analyzed to include progression.
Features of tasks that vary across levels	<ul style="list-style-type: none"> <li>● Familiarity of questions to be answered or problems to be solved?</li> <li>● Sources provided or to be searched by students?</li> <li>● Number of sources to deal with?</li> <li>● Linguistic complexity of sources?</li> <li>● Relevance of the sources in answering questions or solving problems?</li> </ul>	ELP Standards
Features of tasks specific to certain levels	Requirements for students to evaluate the credibility of sources? (Grade 8 only)	ELP Standards

Note. ELP = English language proficiency; KSAs = knowledge, skills, and abilities.

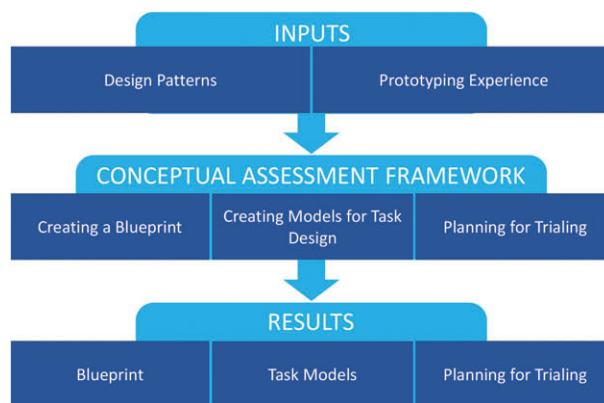


**Table 5** “Communicating Findings” Design Pattern

Attribute	Values	Comments
Title	Communicating findings	Title comes from the ELP Standard 6-8.5. “Communicating findings” is a Level 1 function that forms the basis for the more complex ELP Standard 6-8.5.
Summary	This design pattern is intended to describe tasks assessing test takers’ ability to analyze information in a way that would help them to communicate in order to answer questions or solve problems.	ELP Standard 6-8.5 is “conducts research and evaluates and communicates findings to answer questions or solve problems.”
Rationale	A lot of information is accessible through print and digital sources, but an individual must have the ability to make sense of it and communicate it. In academic tasks, students have to analyze the information they gather in order to communicate answers to questions and solutions to problems.	Our review of the literature and observation of middle school classes revealed the role of language abilities for making sense of information accessed through the use of technology. The construct framework makes these middle school tasks essential because it includes academic contexts as necessary for specifying construct meaning.
Focal KSAs	Labeling collected information Vocabulary of the academic domain Category labels more abstract than the vocabulary of the elements.	The construct framework theorizes functions as the primary category for conceptualizing knowledge, skills, and abilities. Particular functions are called upon by certain tasks and in turn require specific semiotic resources (including both linguistic and nonlinguistic).
Additional KSA	<ul style="list-style-type: none"> <li>● Familiarity with task</li> <li>● Subject-area knowledge</li> <li>● Nonlinguistic semiotic resources for interacting with the technology</li> </ul>	The construct framework theorizes knowledge, skills, and abilities as working together to accomplish the functions called upon by tasks in particular contexts. Test takers’ task familiarity and knowledge inevitably come into play in successful task completion.
Potential observations for evaluation (work products)	Labels selected or created to describe collected information	Work products can be test taker’s selected or produced labels for groups of like elements, whose vocabulary falls within an academic domain.
Potential observations for evaluation (implications for potential rubrics)	Correctness of labels of collected information	Correctness of labeling needs to be specified unambiguously even though academic tasks may be open to multiple labels for certain groups.
Characteristic features of tasks	Tasks requiring students to analyze the characteristics of elements in a set and label them. The task needs to draw upon vocabulary and concepts from an academic area relevant to the scenario. The task needs to be interesting and appear relevant to students so that they will be motivated to engage with the problem.	The task should be presented to appear like the tablet-based tasks that the students perform in their science classes. The software should include gamelike features to help maintain interest and create incentives.
Features of tasks that vary across levels	<ul style="list-style-type: none"> <li>● The familiarity of questions answered or problems solved</li> <li>● Amount of collected information to communicate</li> <li>● Complexity of collected information to communicate</li> </ul>	These features are hypothesized, but will need to be investigated.
Features of tasks specific to certain levels	<ul style="list-style-type: none"> <li>● Using appropriate graphics such as diagrams required? (Levels 3 – 5 only)</li> <li>● Citing sources required? (Levels 3 – 5 only)</li> <li>● Using a standard citation format required? (Levels 4 – 5 only)</li> </ul>	These are based on the ELP Standards.

*Note.* ELP = English language proficiency; KSA = knowledge, skills, and abilities.





**Figure 8** Schematic diagram of inputs, activities, and results for the conceptual assessment framework.

patterns, we sketched the design for a prototype assessment, which provided essential experience to interpret the results of the domain analysis. The nonsequential use of ECD layers is noted by Mislevy and Haertel (2006), who pointed out, “cycles of iteration and refinement both within and across layers are expected and appropriate” (p. 7). Accordingly, it should be noted that the sequential ordering of the domain modeling and CAF steps depicted by the linear organization of this report does not reflect the interactive process that was actually undertaken.

### Conceptual Assessment Framework

The CAF development is the stage of the ECD process where a blueprint is created by considering the original mandate along with the constraints and logistics of the project and the results of the previous steps. The blueprint for the assessment requires a description of “(a) the creation of tasks, evaluation procedures, and statistical models, (b) delivery and operation of the assessment, and (c) analysis of data coming back from the field” (Mislevy, 2011, p. 14). For the Tablet Project, this stage moved seamlessly from the domain modeling phase because it remained centered around developing prototype tasks. During this phase, our prototyping took on additional dimensions to specify the task and evidence models we would need to produce meaningful scores as well as plans for gathering initial data from prospective test users. Figure 8 shows the inputs, CAF activities, and results of this stage and emphasizes the concrete development activities, even though these were affected by other factors in the project as well.

#### Inputs to the Conceptual Assessment Framework Development

The CAF development is undertaken through the use of the design patterns in addition to the experience in prototype development that occurred during the development of the design patterns. These are the two inputs shown in Figure 8. The design patterns provided the detail required for developing student, task, and evidence models by specifying the relevant detail about what should be measured, what the tasks should look like, and the rationales for these specifications. Despite the content of the design patterns, we found that an additional input was essential to move forward with the CAF activities: the experience the team had gained in sketching the design for the prototype task. The concrete experience of designing prototype tasks helped in both establishing the design patterns and in engaging in the CAF development. Fundamental to development of such an expansive set of plans was identification of a content area and potential scenario that would provide the basis of the assessment across the levels.

#### Conceptual Assessment Framework Development Activities

The CAF development activities consisted of iterative, cyclical processes of sketching task-level details and planning macrolevel assessment systems. The goal for creating the task-level detail, as specified in ECD, is to produce three types of models: a student model, a task model, and an evidence model. These three models require precise specifications, which we drafted based on the task design patterns and our experience with the prototype task development to assess the abilities required for ELP Standard 5. Even though we had opted to focus on ELP Standard 5, it was necessary to create

the blueprint for the entire assessment that would span the 10 levels of the ELP Standards to provide materials for learning and assessment at each of the levels. Many of the specific features of the prototype task depended on how it fit within this larger system, whose construction was beyond the scope of this project. For the prototype design, decisions were made in view of the opportunities that would be available for piloting the prototype assessment in classrooms that we would be likely to have access to.

### ***Creating a Blueprint***

The blueprint was created to show the progression students would make as they worked at each of the five levels of development to achieve the competency described in ELP Standard 5: conduct research and evaluate and communicate findings to answer questions. To create the blueprint, we developed parallel designs for the learning and assessment tasks for each of the five levels. This blueprint allowed us to select not only a standard, but also a specific level of the standard for development of the CAF models without being distracted by concerns about the other four levels of the standard. Having selected to develop the CAF models for the first level of the progressions, we were able to create models.

### ***Creating Models for Task Design***

The guidance from ECD about how to create the three models is intended to be applicable across a variety of test development needs, leaving specifics of model development to the task designers. Mislevy and Haertel (2006) described the student model as expressing “what the assessment designer is trying to measure in terms of variables that reflect aspects of students’ proficiencies. Their number, character and granularity are determined to serve the purpose of the assessment” (p. 10). For the tablet assessment, the student model needed to support a total score that was interpretable in terms of ELP Standard 5. The score could reflect the standard holistically, or it could refer to an aspect of the standard as defined by the test developers. The linguistic resources required for ELP Standard 5 were not specified in the ELP Standards, but resources required for three other standards (ELP Standards 8–10) were given, so we examined the way that the progression was described through Levels 1 through 5 for those standards. We then used the semiotic functions from the framework for multimodal communication to elaborate the abilities underlying performance required by that standard.

Mislevy and Haertel (2006) characterized the task model as a way of describing “the environment in which students say, do, or make something to provide evidence. A key design decision is specifying the form in which students’ performances will be captured” (p. 10). Specification of our task model required making decisions about how to use Attributes 6 through 10 from the design patterns to develop more task features. The process got underway when a potential topic and scenario had been drafted, and the assessment series could then be viewed as episodes of a scenario concerned with learning and assessment of a particular standard for a range of middle school grade levels. Based on the scenario, it became possible to specify what kind of performance, or work product, would be elicited from the students.

The work product needs to be decided upon to create the evidence model, which consists of evaluation and measurement components and which connects the student and task models. Mislevy and Haertel (2006) described the first component as follows: “The evaluation component says how one identifies and evaluates the salient aspects of student work, in terms of values of observable variables” (p. 10). For the tablet assessment, Attribute 9 of the design patterns provided the guidance for the evidence model. The authors described the measurement component as being responsible for synthesizing the data that comes from the evaluation component. For the tablet assessment, work products were planned to provide the evidence needed for evaluation, and for each score that was needed, multiple work products were included to allow the measurement component to generate a reliable synthesis of performance.

### ***Planning for Trialing***

The student, task, and evidence models needed to be tested by having real students trial the assessments. We designed a usability study to examine how middle school ELLs would interact with the tablet-based prototype assessment and how the resulting data would reflect the models. We had to identify schools where ELLs were in science classes that would be studying the topics covered in the assessment. We wanted to explore the usability and authenticity of the tasks and items to discover if evidence suggested that task design (i.e., the LOA components in Figure 6) actually promoted the learners’ positive perceptions and use of the assessment as hypothesized in the theory of action (Figure 6). We also wanted to

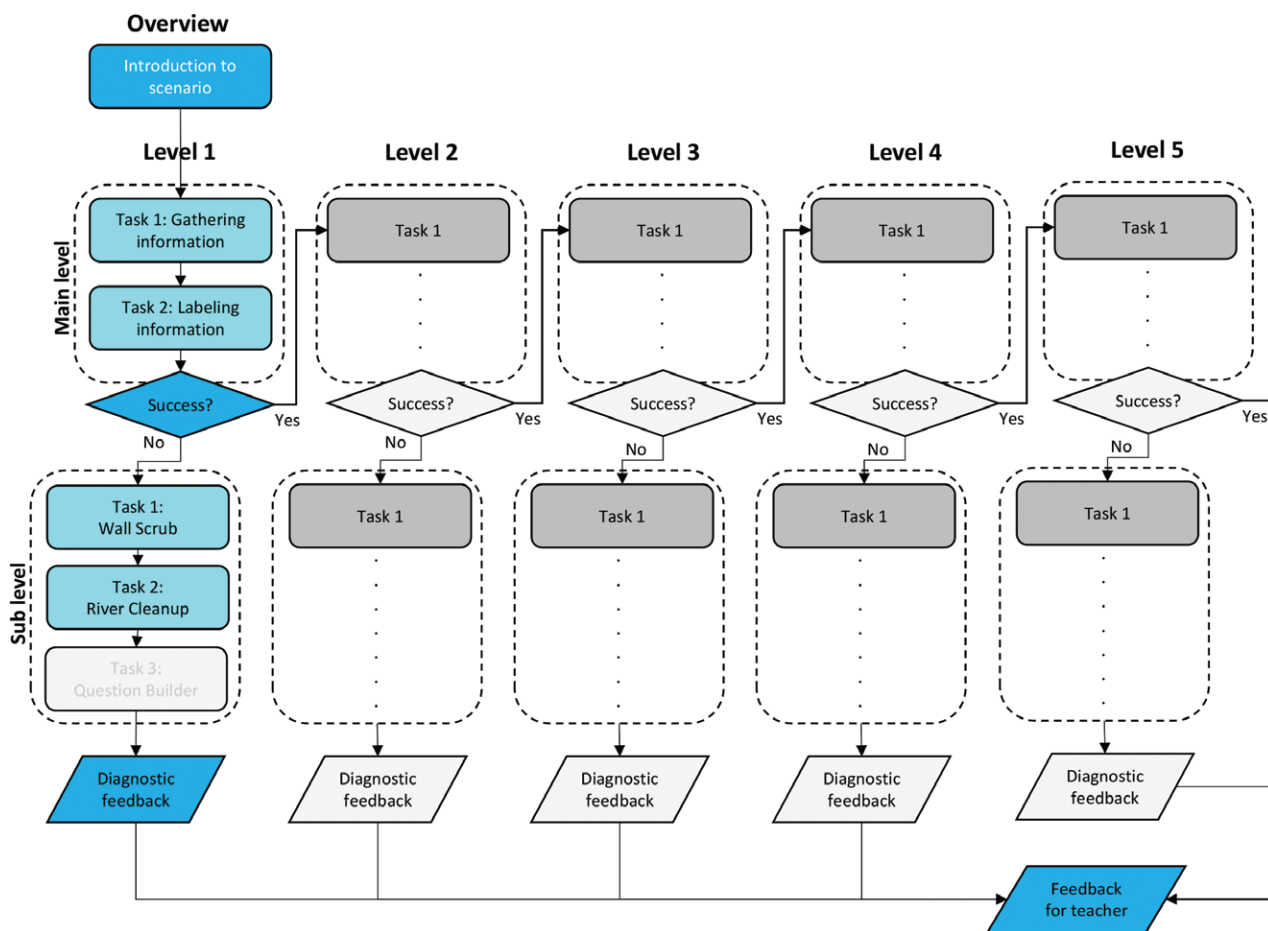


Figure 9 Schematic diagram of the blueprint for the ELP Standard 5 assessment with the scope of the prototype assessment marked in blue.

investigate whether the assessment tasks would elicit the intended work products from students to reveal evidence about developmental proficiency and linguistic resources.

### Results of Conceptual Assessment Framework Development

The results consisted of the blueprint for a complete assessment for all five levels of progression toward ELP Standard 5: the student, evidence, and task models underlying the prototype as well as plans for trialing the test to gather data. The results from the CAF development activities also resulted in a prototype assessment, for which we were able to specify the student, task, and evidence models. We also sketched plans for delivery of the assessment to small numbers of students and showed it to some teachers in order to obtain feedback.

### Blueprint

The schematic diagram for the overall blueprint is shown in Figure 9. It consists of assessment/learning tasks at each of five levels of development for ELP Standard 5. For each of the five levels, the assessment task structure is the same, consisting of main level and sublevel tasks. For the prototype assessment, the main level tasks are designed to provide students the opportunity to use the language functions described by ELP Standard 5 at each developmental level. Based on their success, they either progress to the learning task at the next developmental level (e.g., from main Level 1 to main Level 2) or move to a series of diagnostic assessments at the sublevel.

The five levels of assessment are designed to be tailored to individual students at each of the five levels. An individual student’s path through the assessment—from the overview to receiving diagnostic feedback—will vary based on their

performance on main level tasks, and typically involve the completion of diagnostic tasks. However, regardless of the individual path taken through the assessment, components should be designed to keep students active and engaged for a fixed period. For lower proficiency students, there will be less engagement with the overall narrative through main level tasks and more engagement through the gamelike diagnostic, or sublevel, tasks. Higher proficiency students will spend more time on main level tasks and progressing through the overarching narrative. By adapting the path of the assessment according to students' responses, any path through the assessment should take approximately 40 minutes to complete. The timing is intended to be appropriate for completion during one 50-minute classroom period.

### **Assessment Models**

The results of the CAF development included the student models, task models, and evidence models specified for Level 1 of the assessments for ELP Standard 5, the section highlighted in Figure 9. We analyzed ELP Standard 5 as consisting of two distinct parts: finding information and communicating information. A score should be produced for each of the two parts to be useful to teachers and students for feedback to inform instruction.

A two-part student model was needed in order to report two scores for the overall standard, one indicating the ability to find information and the other indicating the ability to communicate information. In addition, for students who did not score at a certain level, each model contained information at a more granular level to produce scores to reflect the underlying linguistic resources that would contribute to performance on the overall task. The main level student models are defined with reference to the functions stated in the ELP Standards, and the sub level information in these models is specified with reference to the multimodal framework for communication and learning.

The task models specify the episodes in the overall narrative of the assessment and the manner in which the students interact with the narrative to produce their work products. Expressing the LOA components in the theory of action (Figure 6), Attributes 6–10 of the design patterns were intended to increase students' motivation to engage with the task. Motivation is promoted by the task features that shape the form of the interaction with the test takers. First, feedback was provided in two forms: (a) as instantaneous results from students' actions in the form of points and stars (visual), dings, and buzzes and (b) as delayed indicators of achievement through medals to summarize task performance (gold, silver, bronze). Second, incentives were provided by allowing students to acquire points and medals. There were also aesthetic incentives, such as uncovering a mural in Diagnostic Task 1 or changing the color of the river in a diagnostic task. Third, the environment was made attractive through a comic book, film noir art aesthetic; use of video and audio in sources; use of an overall narrative; and customizable detective avatars.

The scenario entails a macrolevel goal intended to capture students' motivation and interest by engaging them in a problem-solving activity that requires them to use language to perform the functions specified in the ELP Standards. In the scenario, the student plays the role of a detective who attempts to figure out and solve an environmental issue. At the lowest level, the student is asked to gather some information and label it before communicating it to others. Doing so requires the use of low-level vocabulary on the topic of the environment, and the tasks are designed to require microlevel skills to accomplish tasks leading to the goal. If the tasks are sufficiently clear and the students sufficiently engaged to want to perform well, they should elicit the information needed for the student model.

The task model also includes elements of global design principles in support of making the tasks clear and motivating for the students. Clarity is attempted in the way that the task communicates to students. First, all detective thought bubbles should be both text and audio. Users should be able to replay the audio. Second, any process data that can be captured (e.g., time spent on each task and item, number of replays of content, number of taps on various resources) should be captured for research analysis. Linguistic input should be used sparingly, particularly for task directions that can be visually communicated using animation or other nonlinguistic models.

The evidence model was based on Attributes 8–10 of the design patterns. We defined characteristic features of tasks (8), features that can vary (9), and features specific to certain levels (10) as a starting point to determine how many tasks would be needed to represent a variety of task variations and achieve stable results at each level. The target number of items for each of the main tasks was the high end of the 20–30 range for the prototype, with the intent of providing adequate information to make a reliable judgment about the adequacy of performance on the main task. For the diagnostic tasks, we needed even more items because we intended to code diagnostic tasks based on multiple features (e.g., vocabulary; part of speech, Tier 1–3). These considerations at the design stage laid the foundation for large-scale trialing to estimate

item parameters and potentially even computer-adaptive presentation of items for more efficiency. However, before such large-scale trialing could be attempted, we first planned a small-scale usability study.

### Plans for Trialing

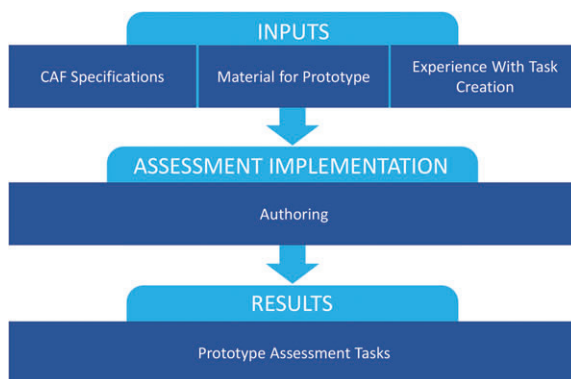
The usability study was planned for small-scale data collection to show how well the task models operated for students at the target level. Specifically, the research questions were developed to investigate how well students were able to understand and navigate the tasks and what kind of evidence the models would capture about test takers' developmental proficiency and their linguistic resources, as well as the students' perceptions of the prototype assessment. We also wanted to observe whether the task models were effective for gathering test takers' performance under appropriate conditions to obtain good samples and whether the design features intended to promote learning (such as feedback) were distorting the quality of the samples. We also wanted to assess the students' perceptions toward the assessment because the design included task features intended to prompt the test takers to see the app as a gamelike and aesthetically pleasing environment and to feel motivated by the avatars and animated agents that were integral to the task designs. Concerned with these fundamental issues of task implementation, the usability study was necessarily small and not intended to provide definitive estimates of item characteristics. Nonetheless, the small-scale usability study was useful because it uncovered sources of confusion for students during task performance. These results are summarized below, after the description of the assessment implementation.

### Assessment Implementation

The assessment implementation stage of ECD is the time for “constructing and preparing the operational elements specified in the CAF” (Mislevy, 2011, p. 16). What this entails therefore depends on what is specified in the particular CAF for the assessment as well as the test developers' means for implementing the assessment, which are largely influenced by the aspects of the mandate, including the timing, funding, and logistics. For the Tablet Project as shown in Figure 10, the primary activity was authoring to instantiate the tasks that had been designed during the creation of the CAF. Software development was contracted to an outside company that was responsible for building software as specified by the planning documents, including the front-end look and feel of the assessment as well as the content and functionality. A large part of the authoring activities, therefore, included an iterative process of communication with these external developers. The results of the assessment implementation phase were the prototype assessments.

### Input to Assessment Implementation

As input to assessment implementation, we developed the CAF specifications and the content for the prototype for Level 1 of the assessment. The draft content and plans consisted of login and avatar selection screens, an introductory sequence, two main level tasks, and two diagnostic tasks measuring linguistic resources expected to be essential for performance on the main level tasks. We also relied on our experience with task creation, which had begun even before finalizing the CAF.



**Figure 10** Inputs, authoring activity, and results for assessment implementation.



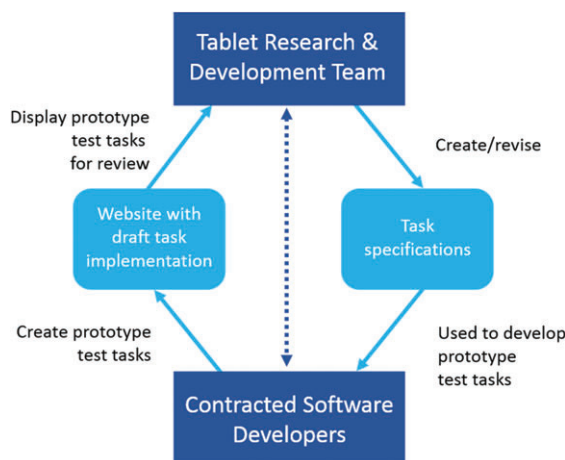


Figure 11 The authoring activities undertaken during the assessment implementation phase.

## Authoring

Authoring refers to the integrated process of software creation, which encompasses both content and functionality in a piece of software. It is the critical phase of the project, where the concepts and structures that have been planned are instantiated in the working assessment. The authoring process in the Tablet Project over a 9-month period was accomplished through a cycle of communication, creation, review, and revision. This process is depicted schematically in Figure 11.

Figure 11 shows the two primary teams that participated in authoring: our research and development team (top) and the contracted software developers (bottom). The multifaceted communication patterns between the two teams are shown by the arrows. The primary artifacts supporting communication are shown on the right and left sides as the “task specifications” and the “website with draft task implementation,” respectively. The research team communicated their plans for the prototype task by providing task specifications, which the software developers used to create prototype tasks. The prototype tasks were accessible to the research and development team on a website created by the software developers. The research and development team reviewed the draft tasks and requested revisions when necessary and also modified the specifications as needed.

Because the Tablet Project was intended to explore the capabilities of the technology, the initial specifications included innovative elements, which were proposed without knowledge of the cost of implementation. The software developers then had to provide input into the feasibility of some of the aspects of the design, resulting in certain compromises being made in the specifications. These compromises were negotiated with the developer as they arose, and the prototype task description document was edited and updated as task designs were revised. For example, the ETS design team’s original vision for the vocabulary diagnostic task was much more complex than what was executed in the prototype and involved a large amount of animation, panning, and zooming to depict a growing tree. This design proved to be too demanding on the project’s resources, and the current vocabulary diagnostic task was designed as a compromise in close collaboration with the developer.

In Figure 11, the dotted-line, double-headed arrow in the middle represents the regular communication that occurred between the development team and the software developers, both in writing and in weekly conference calls. This authoring process supported the multiple rounds of feedback and revision the assessment tasks went through before they were finalized. While the draft content and functionality was being implemented, the construct framework and prototype assessment material (source texts, items) continued to be reviewed for quality and fairness.

## Results: Prototype Assessment Tasks

The prototype assessment tasks require test takers to assume the role of a detective who is working to solve problems in his or her town, Citytown. We elaborated an overarching narrative that was intended to be accessible and engage with the





Figure 12 A beginning screen of the assessment that invites students to create their avatar.

detective and investigation concept. The concept would accommodate five main level tasks, constructed based on both the task patterns and a relevant topic from middle school science standards, photosynthesis:

Citytown is interested in entering a contest for the “Best Town.” The winning town needs to be an environmentally friendly and pleasant place to live. It needs to have lots of trees, green parks, clean streets, and happy and healthy people and animals. The people of Citytown really want to win this contest, but are currently facing some environmental problems. Citytown and its residents have a “green” problem. The trees and the flowers are not growing like they used to, and the parks are no longer green and filled with flowers, trees, and animals. There is more pollution and trash in the streets, and energy resources are being wasted. Citytown needs to do something if they want to win the Best Town Contest!

In order to solve these problems and win the contest, Citytown decides to hire a detective to help them understand how to make Citytown greener. The student plays the role of the detective and uses a special tablet and Noodle glasses to help solve cases.

A mysterious letter is sent to the detective explaining that Citytown has a green problem and invites the detective to figure out in what ways Citytown is green or not green. If the detective is successful, then (s)he is hired to help Citytown become greener in order to win the Best Town Contest. The mayor asks the detective to look into the trees in Citytown Park and see why they are not growing well. This guides the detective to learn more about photosynthesis at each level to understand where trees and plants get their energy. (Schmidgall, Lopez, Blood, & Wain, 2015, pp. 109–110)

The initial prototype included three main parts: an overview, Level 1 main tasks, and Level 1 diagnostic tasks. Because all students were expected to understand the overview sequences regardless of their proficiency level, we emphasized the use of visuals to convey information. The overarching narrative described above was translated to a comic booklike sequence that used short, simple sentences with high-frequency vocabulary. All written text was accompanied by audio.

The overview consists of the login, avatar selection, and introductory sequences. When students open the app, they log in, choose a nickname, and customize their detective avatar (gender; and skin, hair, and eye color; see Figure 12). Students select a case to investigate (currently, only The Green Problem is available), and an introductory sequence provides background on the detective character, the functionality of the interface, and an overview of the narrative. The purpose of this section is to orient and interest the students in the task but not to make any assessment of ability.

The second part consists of two main level tasks. In main level Task 1, test takers view an introductory sequence that presents the focal problem for Level 1: “What makes a town green?” Test takers then view 30 short (20–50 word) text and video sources presented as extracts from articles and online videos. Figure 13 shows a screen in this main level task, where

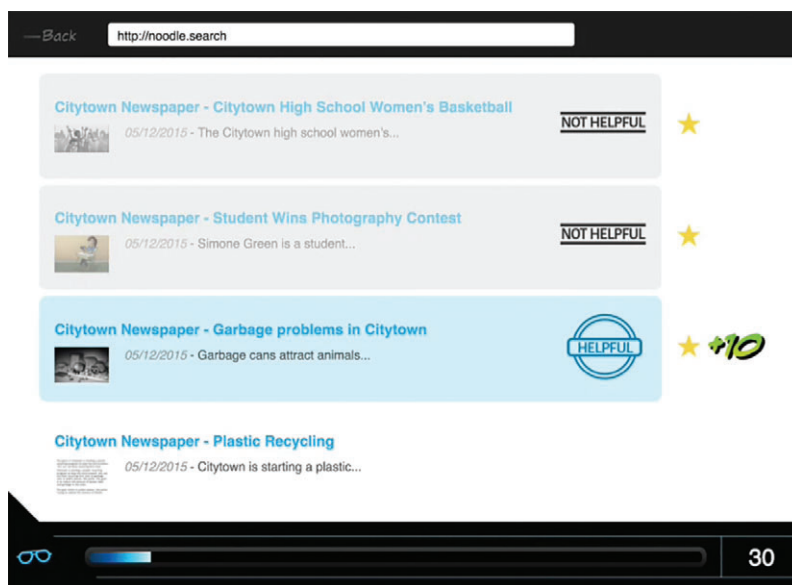


Figure 13 A screen showing main level Task 1 requiring students to categorize information sources as *helpful* or *not helpful* for the problem-solving task.

students have categorized sources as *helpful* or *not helpful* to indicate whether they contain information that is relevant to Citytown’s green problem. In the detective problem-solving scenario, the helpful sources are clues. Upon categorizing each source, students receive immediate feedback on whether their response was correct or incorrect and a reward for a correct response.

Feedback is provided aurally (“ding” for correct, “buzz” for incorrect) and visually. As shown in Figure 13, visual feedback includes the presence or absence of a star next to each source and a stamp next to the source indicating whether it is helpful or not helpful. This immediate feedback is intended to promote self-regulation and self-efficacy and facilitate uptake. After students categorize each source once, they are required to reattempt sources that were categorized incorrectly. This feature of the task was intended to reinforce the importance of the immediate feedback and provide a mechanism for measuring uptake.

Rewards are provided in the form of points that accumulate in the student’s total score in the lower right corner of the screen. Ten points are awarded for each correct response during the first attempt to categorize sources, and five points are awarded for correct responses during the second attempt. The use of points as rewards is intended to enhance learner self-efficacy, self-regulation, and learner interest and engagement.

In the second main level task, test takers view each of the 20 helpful sources (clues) again. This time they indicate whether each source (clue) *helps environment* or *hurts environment*. Figure 14 shows one of the screens that presents one of the sources. Students need to comprehend this clue and then label it as *hurts environment*. Feedback (aural and visual) and rewards (points) are implemented in the same manner as described for Task 1. As in Task 1, students can reattempt to categorize sources incorrectly labeled on the first attempt. It is intended to assess the students’ use of the semiotic function of using information as well as utilizing multimodal reading and listening comprehension skills.

According to the test blueprint (see Figure 9), the second part of the prototype assessment consists of two diagnostic tasks designed for test takers who cannot complete the Level 1 main tasks successfully. The tasks are contextualized by a transitional sequence that guides them to Citytown Park, where they encounter pollution in several forms. The first diagnostic task focuses on assessing vocabulary knowledge in a gamelike scenario. The student detective begins by discovering a neglected wall in the park and learns that by answering items correctly, they can solve the problem of the dirty wall. In each item in the task, test takers view four words—a stem and three answer choices. They must choose the answer choice that is closest in meaning to the stem. The words are all taken from Marazano and Simms’ (2013) *Vocabulary for the Common Core* and have been coded by domain (everyday, general academic, academic-specific), as well as part of speech (noun, verb, adjective). In this case, the academic-specific domain pertained to middle school science.



**Figure 14** A screen showing main level Task 2 requiring students to categorize information sources as *helps environment* or *hurts environment* part of the problem-solving task.

As with the main level tasks, the sublevel diagnostic tasks include immediate feedback and rewards that are intended to enhance self-efficacy, self-regulation, and engagement. For the first diagnostic task, immediate feedback is provided aurally (ding for correct, buzz for incorrect). For each item, test takers are given two attempts to choose the correct word; if an incorrect answer is chosen on the first attempt, the selected answer choice disappears and negative feedback is provided. As words are correctly chosen, students are rewarded with points and are able to scrub dirt off a wall in Citytown Park, revealing a mural, as shown in Figure 15. Test takers earn points based on each item answered correctly on the first attempt (full points) or second attempt (partial points) and are rewarded with the ability to scrub a portion of a dirty wall with each correct response. As test takers scrub the wall, they begin to reveal a colorful mural underneath as an additional reward. Items are expected to be progressively more difficult based on word frequency estimates and are delivered in three sets of 30 items each that correspond to three separate murals. Thus, the activity is gamelike in that it directs students toward a larger goal connected to the overarching narrative (i.e., making Citytown Park greener), gives them an engaging mechanism to achieve the current goal of cleaning the wall (i.e., the ability to choose which sections to scrub), and offers incremental rewards through the gradual reveal of the mural.

In the second diagnostic task, test takers view three multiword sequences that are slowly floating down Citytown River. The word sequences appear above pieces of trash that must be cleared from the river by correctly arranging them as a simple sentence (subject, verb, object). For each item, test takers swipe back and forth to reorder the chunks, and when they believe the words have been reordered correctly, they tap a garbage can to attempt to remove them, as shown in Figure 16. Each response to an item is given immediate aural feedback (ding for correct, buzz for incorrect). As test takers correctly order sentences, they are rewarded with points and trash is removed from the river; consequently, the river begins to turn from brown to blue, fish return to the river, and plants and animals begin to return to the riverbank. Again, test takers are given the opportunity to reattempt to answer each item after the first incorrect attempt for partial credit. Incorrect second attempts lead to the accumulation of trash under a bridge. Test takers attempt 20 items. This diagnostic task is intended to assess the test taker's ability to order subject, object, and verb chunks for present simple and present continuous tenses and singular and plural subjects.

## Next Steps

The implementation phase of the Tablet Project entailed an interactive process of authoring that resulted in a prototype assessment to be used in trialing that would feed back into at least one more round of revisions before it would be ready for the next stage of ECD, assessment delivery. Assessment delivery encompasses the process of students interacting



Figure 15 A screen showing the reward page of a diagnostic level task requiring students to select a synonym for a content-relevant word.

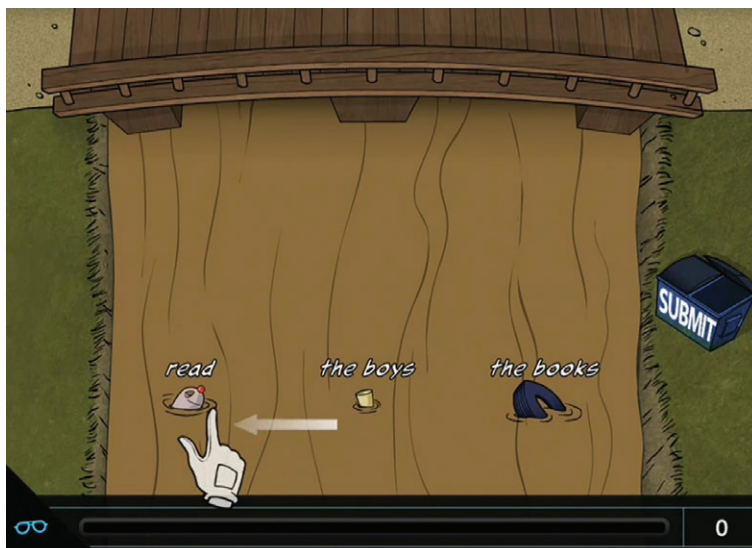


Figure 16 A screen showing a diagnostic level task requiring students to order phrases into sentences.

with the assessment while their work products are being captured and evaluated by the evidence model. The evidence model, in turn, provides data needed to summarize students’ performance in reports and adapt the path through the assessment to select tasks appropriately. The scope of the project did not include the delivery phase, but it did allow for an initial round of trialing in the form of usability research and focus groups to provide input to an iterative design process. The usability testing was intended to provide an initial indication of how prospective users would respond to the features that had been designed to promote their engagement and result in evidence of the defined language abilities.

One of the features in the original design that was not implemented in the authoring was the posttask summative and diagnostic feedback. We therefore created visual mock-ups to conceptualize potential feedback options for learners and teachers. The purpose of these mock-ups was to indicate the type of feedback that could potentially be generated from learner performance and obtain feedback from stakeholders (e.g., teachers, learners) on the attractiveness and perceived usefulness of potential feedback. Although establishing an underlying scoring model would be a critical step in backing assumptions about the meaning of this feedback, this was not a focus for this preliminary stage. Mock-ups of potential feedback to learners are shown in Figure 17.





Figure 17 Mock-ups of end-of-task (left) and end-of-level (right) summative feedback for learners.



Figure 18 Mock-ups of detailed feedback for learners.

In addition to the instantaneous feedback provided to learners for each item within a task, we intended to provide summative feedback after each task and level. The left side of Figure 17 shows an example of how feedback and a reward could be simultaneously provided upon completion of a main level task. After completing the gathering information main level task, a learner would be given feedback on their overall performance (e.g., based on points) in the form of a gold, silver, or bronze medal for their investigation skills. In this example, the learner obtained a silver medal for *investigation* based on obtaining 180 points in the gathering information task. Upon completion of the second main level task, learners would likewise be awarded a medal for *using clues*.

As discussed in the previous section, the design of the app called for learners to either move to the next level or complete a series of diagnostic tasks after completing the two main level tasks. The right side of Figure 17 demonstrates one way in which feedback may be provided to learners at this point. Although the scoring model had not yet been established, it was presumed that successful performance (e.g., meeting a threshold level of points across tasks, or all gold medals) could be indicated by awarding a badge designed for that level—for example, a *junior detective* badge for the successful completion of Level 1. Thus, each badge corresponds to a developmental proficiency level in the standards.

The design of the app also called for learners to complete a series of activities based on their performance. Unless learners completed all five main levels successfully—demonstrating full proficiency with respect to the overall standard—they would be expected to complete a series of sublevel or diagnostic tasks. A visual mock-up of what the more detailed diagnostic feedback might look like for learners at Level 1 is shown in Figure 18.

In the prototype app, the diagnostic tasks for Level 1 included the Wall Scrub and River Cleanup. As shown in the mock-up in the upper corner, summative feedback may be provided in the form of medals, as in the main level tasks. By tapping on one of the tasks, the learner could prompt the app to provide more detailed information about their performance, as

<span style="color: green;">●</span> Full Mastery <span style="color: yellow;">●</span> Int. Mastery <span style="color: red;">●</span> Novice Mastery		
Name	Level	Diagnostic
Bolivar, Margarita	2	<span style="color: red;">●</span> Grammatical Knowledge <span style="color: green;">●</span> Vocabulary Knowledge <span style="color: yellow;">●</span> Discourse Knowledge
Castro, Eric	2	<span style="color: yellow;">●</span> Grammatical Knowledge <span style="color: green;">●</span> Vocabulary Knowledge <span style="color: yellow;">●</span> Discourse Knowledge
Jolibois, Michel	1	<span style="color: red;">●</span> Grammatical Knowledge <span style="color: red;">●</span> Vocabulary Knowledge <span style="color: yellow;">●</span> Discourse Knowledge
Patel, Dev	3	<span style="color: green;">●</span> Grammatical Knowledge <span style="color: green;">●</span> Vocabulary Knowledge <span style="color: yellow;">●</span> Discourse Knowledge

Figure 19 Mock-up of general feedback for teachers.

shown in the lower corner. In this example mock-up, each row corresponds to an item from the Wall Scrub task. The target vocabulary word for each item begins each row. The double-sided green arrow connects the target word to the most related word among the three options presented. The most related word among the options appears in a shade of green, and the unrelated words appear in shades of red. Words selected by the learner during the task appear darker, and the points obtained for each item appear at the end of each row. By tapping the arrows connecting the two words, the learner can obtain text that explains their relationship.

The app was also designed to provide feedback to teachers to support instructional decisions, such as what to teach next, how to group learners for differentiated instruction, or how to formulate individualized teaching plans. Figure 19 shows an example of how more general feedback may be provided to teachers.

Figure 19 shows how feedback would be provided for each individual in a group of learners who completed the tasks. The *Level* indicates the learner's assessed developmental level (1 to 5) with respect to the ELP Standard. This is the most general feedback that is offered, but a teacher may want to know more about the language knowledge and skills that may be holding the student back from achieving a mastery score at the next developmental level. In this mock-up, basic diagnostic information is provided in the form of an evaluation of the degree of mastery demonstrated (novice, intermediate, full) with respect to foundational aspects of language knowledge (grammatical, vocabulary, discourse). For example, Margarita Bolivar has demonstrated that she can successfully complete tasks associated with the first developmental level, but has shown novice mastery of the grammatical knowledge associated with the second developmental level of the standards.

As shown in the mock-up in Figure 20, finer-grained diagnostic information could potentially be provided for individual learners (or groups of learners) based on the intended design. In this example, Michel Jolibois' overall vocabulary strength was assessed at the novice level with respect to the next development level he needs to achieve. This is consistent with the information provided about vocabulary knowledge in Figure 19. This overall evaluation of vocabulary knowledge is further parsed into relevant components such as everyday vocabulary, general academic vocabulary, specialized academic vocabulary, and parts of speech (nouns, adjectives, verbs). The results would be interpreted to indicate that this learner needs assistance with general and specialized academic vocabulary, but his everyday vocabulary is strong.

These additional materials were used along with the prototype tasks to obtain feedback from students and teachers in the usability study and teacher focus groups, respectively, which are described in the next section as they pertain to the first draft of the validity argument for interpretation and use of the assessment.

## Validity Argument

The overall schema for test design and validation (Figure 1) shows that the immediate use for the frameworks, concepts, and prototype tasks created during the test design phase of the project is to provide the detail required to develop a validity argument for test interpretation and use. Using Kane's (2006, 2013) conception of validity argument as a means for identifying the evidence needed to support interpretations and uses of scores from assessments, we began by first sketching an interpretation/use argument, which makes explicit the intended interpretations of the assessment results and their intended use. The use in our case is multifaceted because the assessment is intended as an LOA, which encompasses



Jolibois, Michel (Level 1): Vocabulary Diagnostics	
Overall Vocabulary Strength:	● Novice Mastery
Everyday Vocabulary Strength:	● Full Mastery
General Academic Words Strength:	● Novice Mastery
Specialized Academic Words (Math, Science):	● Novice Mastery
Noun Strength:	● Int. Mastery
Adjective Strength:	● Int. Mastery
Verb Strength:	● Novice Mastery

Figure 20 Mock-up of detailed feedback for teachers.

both measurement and learning goals. In the domain analysis, we therefore created not only the construct framework that would provide a basis for stating measurement outcomes of the assessment but also a theory of action framework that allows us to make explicit the connection between assessment design and learning outcomes. Both the measurement and learning facets of the assessment are represented in the interpretation/use argument for the tablet assessment. With the assumptions in the interpretation/use argument made explicit, we interpret the findings from the usability study and teacher focus groups with respect to their support for particular assumptions stated in the argument.

### Interpretation/Use Argument

The interpretation/use argument for the tablet-based assessment specifies the intended interpretations for the assessment results, their uses in an educational context, and their consequences for those involved in the assessment. An interpretation/use argument is intended to make explicit the assessment developers' intentions for an assessment in a format that can serve as a plan for validation research (Kane, 1992, 2001, 2013). In an ideal world, the development would begin with the test designers' vision for creating tests that contribute to making a better world. In the real practices of test designers, such altruistic motives may underlie test development, but typically specific testing projects like the Tablet Project are guided by a mandate that may include with more or less specificity the intended use, which may imply some consequences. To develop the interpretation/use argument for the tablet assessment, we therefore began with the mandate, which specified several of the critical elements for the argument, as follows:

In the Tablet Project the mandate specified that the test was to assess *academic language performance* in a tablet-based environment to correspond with the way that tablets would be used in the classroom for communication and learning in middle schools of the future. The tests were intended to be used in *formative classroom assessment*, the purpose of which was to provide relevant detailed feedback to teachers and students in addition to total scores whose meaning is relevant to classroom instruction at that level. The assessments were intended to gather data that would serve in making measurement-based inferences about students' abilities in addition to *creating opportunities for students to learn*.

In the above mandate, three critical components that need to appear in the interpretation/use argument are included. *Academic language performance* indicates the construct intended to underlie score interpretation. *Formative classroom assessment* indicates that the assessment results should be informative to students and teachers during the learning process. The fact that the assessment should *create opportunities for students to learn* means that the assessment results should have a broader impact on the learning context. Each of these components was developed theoretically to provide the necessary detail for assessment design and validation. The resulting frameworks were used in the ECD process, and the interpretation/use argument sets the stage for them to be used in validation as well.

The interpretation/use argument, shown in Figure 21, includes the assessment construct in the explanation inference, the formative test use in the effectiveness inference, and the learning opportunities in the learning inference. In addition, it includes inferences that make explicit other intended meanings that are implicit in a general mandate. Each inference

denotes a particular type of meaning that the scores are intended to have when the assessment is used. For example, Inference C in Figure 21, generalization, indicates that the scores are intended to be stable across similar tasks in a defined domain of assessment tasks. The basic meaning of the measurement-related inferences comes from Kane's presentation of validity argument and from their use in other language assessment validity arguments (e.g., Chapelle, Enright, & Jamieson, 2008). Each refers to an inference that is made when the assessment results are used about the quality of the domain analysis (domain definition), performance elicitation and scoring processes (evaluation), score consistency (generalization), reflection of the intended construct (explanation), relevance to the intended domain (extrapolation), and appropriateness of use (utilization). The learning-related inferences were developed to reflect the quality of the consequences of assessment for teachers and students in terms of effectiveness and specifically for learning as specified in the theory of action. The specific meanings of inferences in the interpretation/use argument are given in the warrants and assumptions underlying each one.

The overall interpretation/use argument is read from the bottom to the top to symbolize the intent of having each of the sets of inferences, warrants, and assumptions serve as grounds upon which the next level of inference, warrant(s), and assumptions builds. In Figure 21, the inferences with their corresponding warrants and assumptions are designated with a letter (beginning with A at the bottom) and progressing sequentially to G at the top. The grounding for the entire interpretation/use argument is the domain definition inference. In order for users of the assessment to be justified in making the domain definition inference, the warrant in need of support is that observations of performance on the tablet assessment reveal relevant knowledge, skills, processes, and strategies representative of those required for app-based tasks in the middle school classroom. The assumptions underlying this warrant are (a) that the domain analysis reveals adequate knowledge about communication and learning in middle school content classes to create useful task patterns, a construct definition, and a theory of learning as well as to uncover valued conceptions of learning progressions, and (b) that task patterns are adequate to serve as a basis for developing tasks reflective of app-based tasks in the middle school classroom. Support, or backing, for these assumptions comes from the domain analysis and domain modeling processes of ECD. The inclusion of a domain definition inference in the overall interpretation/use argument provides a place for the ECD process in validation: All aspects of validation rest on the quality of the test development process.

If there is adequate support for the assumptions underlying the domain definition inference, its conclusion serves as grounds for the next inference, evaluation (B in Figure 21). The evaluation inference requires that scores and feedback provide teachers and students with accurate information. This warrant in turn requires support for these assumptions:

- 1 Performance is gathered under appropriate conditions to obtain good samples.
- 2 Design features intended to promote learning do not distort the quality of the samples.
- 3 Scoring rules in the evidence model are designed and implemented to provide relevant scores.
- 4 Feedback reflects the outcomes of the scoring rules.
- 5 Feedback is perceived as accurate to experts.

These assumptions about obtaining relevant performance from test takers and scoring it accurately depend on processes put into place during test development. They can begin to be examined through usability testing and eliciting expert judgment, as shown in the description of the usability and teacher focus group studies below.

If usability testing and expert judgment from focus groups support the assumptions underlying the evaluation inference, its conclusion provides grounds for the generalization inference, which has two warrants. One is that scores are consistent across tasks, forms, and occasions of assessment; in other words, that its results are reliable. This warrant requires backing for three assumptions recognizable as essential aspects of reliability: (a) tasks are designed according to consistent specifications, (b) scores are free from unmotivated sources of variation, and (c) tasks used for computing scores are sufficient in number to capture performance consistency. The second warrant is that feedback is consistent across multiple instances of the same error. This warrant about feedback requires backing for two assumptions: (a) assessments provide sufficient opportunities for students to obtain frequent error feedback for learning, and (b) error recognition is consistent in identifying errors in test takers' performance.

If the generalization inference is supported, the conclusion is that the scores are sufficiently stable to reflect a construct that we should be able to explain, and it therefore provides grounds for the explanation inference. The explanation inference is made on the basis of the warrant that scores and diagnostic results on the test reflect intended aspects of the construct of academic language performance. This warrant is based on three assumptions: (a) students report using the intended linguistic resources during task completion, (b) students are observed to be using intended linguistic resources







Inference	Warrants and Assumptions (listed and numbered under each warrant)
<p>G. Learning</p> 	<p><b>Warrant 1: Students' learning improves with respect to the ELP Standards.</b></p> <ol style="list-style-type: none"> <li>Teaching improves by becoming more efficient, systematic, and responsive to students' needs.</li> <li>Students focus their independent study on learning goals.</li> </ol> <p><b>Warrant 2: Students become more autonomous learners.</b></p> <ol style="list-style-type: none"> <li>Students' self-efficacy is improved by the use of the app.</li> <li>Students' interest and engagement is improved by the use of the app.</li> </ol>
<p>F. Effectiveness</p> 	<p><b>Warrant 1: Teachers target instruction to KSAs that students need to achieve learning goals.</b></p> <ol style="list-style-type: none"> <li>Teachers interpret diagnostic feedback as relevant to developmental standards.</li> <li>Teachers use diagnostic feedback to plan and modify instruction.</li> </ol> <p><b>Warrant 2: Students focus study on abilities they need to acquire to gain mastery of the ELP Standards.</b></p> <ol style="list-style-type: none"> <li>Students obtain relevant diagnostic feedback from the app.</li> <li>Students comprehend and interpret diagnostic feedback as relevant to their learning goals.</li> </ol> <p><b>Warrant 3: Students use the app to explore its features.</b></p> <ol style="list-style-type: none"> <li>Learners obtain item-level and task-level feedback.</li> <li>Learners obtain rewards for completing tasks.</li> <li>Learners perceive the app as a game-like and aesthetically pleasing environment.</li> <li>Learners are immersed in the narrative used as a basis for the app.</li> <li>Learners are motivated by the avatars and animated agents.</li> </ol>
<p>E. Extrapolation</p> 	<p><b>Warrant: Scores and diagnostic results are relevant to performance required for success in middle school content area courses.</b></p> <ol style="list-style-type: none"> <li>Diagnostic results from the assessment are relevant to linguistics resources required for tasks in the content classroom.</li> <li>The characteristics of the assessment tasks correspond to the content tasks required in content courses in middle schools.</li> <li>The criteria and procedures for evaluating the performance on the assessment are pertinent to those identified by instructors as important for assessing performance in class.</li> </ol>
<p>D. Explanation</p> 	<p><b>Warrant: Scores and diagnostic results reflect intended aspects of the construct of academic Language performance.</b></p> <ol style="list-style-type: none"> <li>Students report using the intended linguistic resources during task completion.</li> <li>Students are observed to be using intended linguistic resources during task completion.</li> <li>Scores on the tablet assessment are related in predicted ways to scores on other measures.</li> </ol>
<p>C. Generalization</p> 	<p><b>Warrant 1: Scores are consistent across tasks, forms, and occasions.</b></p> <ol style="list-style-type: none"> <li>Tasks are designed according to consistent specifications.</li> <li>Scores are free from unmotivated sources of variation.</li> <li>Tasks used for computing scores are sufficient in number to capture performance consistency.</li> </ol> <p><b>Warrant 2: Feedback is consistent across multiple instances of the same error.</b></p> <ol style="list-style-type: none"> <li>Assessments provide sufficient opportunities for students to obtain frequent error feedback for learning.</li> <li>Error recognition is consistent in identifying errors in test takers' performance.</li> </ol>
<p>B. Evaluation</p> 	<p><b>Warrant: Scores and feedback provide teachers and students with accurate information.</b></p> <ol style="list-style-type: none"> <li>Performance is gathered under appropriate conditions to obtain good samples.</li> <li>Design features intended to promote learning do not distort the quality of the samples.</li> <li>Scoring rules in the evidence model are designed and implemented to provide relevant scores.</li> <li>Feedback reflects the outcomes of the scoring rules.</li> <li>Feedback is perceived as accurate to experts.</li> </ol>
<p>A. Domain Definition</p>	<p><b>Warrant: Observations of performance on the assessment reveal relevant knowledge, skills, processes, and strategies representative of those required for app-based tasks in the middle school classroom.</b></p> <ol style="list-style-type: none"> <li>Domain analysis revealed adequate knowledge about communication and learning in middle school content classes to create useful task patterns, construct definition, and theory of learning as well as to uncover valued conceptions of learning progressions.</li> <li>Task patterns are adequate to serve as a basis for developing tasks reflective app-based tasks in the middle school classroom.</li> </ol>

Figure 21 Interpretation/use argument for tablet-based English language assessment.



during task completion, and (c) scores on the tablet assessment are related in predicted ways to scores on other measures. Support for these assumptions requires research designed to address specific research questions, the first two using qualitative methodology and the final one using a correlational study.

Support for the explanation inference is needed to draw a conclusion that can be used as grounds for making an extrapolation inference. The extrapolation inference requires support for the warrant that scores and diagnostic results are relevant to performance required for success in middle school content area courses. This warrant is based on three assumptions: (a) diagnostic results from the assessment are relevant to linguistic resources required for tasks in the content classroom, (b) the characteristics of the assessment tasks correspond to the content tasks required in content courses in middle schools, and (c) the criteria and procedures for evaluating the performance on the assessment are pertinent to those identified by instructors as important for assessing performance in class.

If the extrapolation inference is supported, its results serve as grounds for an inference about the effectiveness of the assessment results. An inference about effectiveness requires support for three warrants. The first is that the teachers target instruction to KSAs that students need to achieve learning goals. This warrant is based on two assumptions: (a) teachers interpret diagnostic feedback as relevant to developmental standards and (b) the teachers use diagnostic feedback to plan and modify instruction. The second warrant is that students focus study on the abilities they need to acquire to gain mastery of the standards. This warrant requires support for two assumptions: (a) students obtain relevant diagnostic feedback from the app, and (b) students comprehend and interpret diagnostic feedback as relevant to their learning goals. The third warrant is that students use the app to benefit from its features. The assumptions underlying this warrant are that (a) learners obtain item-level and task-level feedback, (b) learners obtain rewards for completing tasks, (c) learners perceive the app as a gamelike and aesthetically pleasing environment, (d) learners are immersed in the narrative used as a basis for the app, and (e) learners are motivated by the avatars and animated agents.

If the effectiveness inference can be made, its conclusion serves as grounds for the learning inference, whose first warrant states that students' learning improves with respect to the ELP Standards. This warrant requires support for two assumptions: (a) teaching improves by becoming more efficient, systematic, and responsive to students' needs, and (b) students focus their independent study on learning goals as learner autonomy increases. Its second warrant is that students become more autonomous learners. Support for this warrant requires backing for the assumptions that students' self-efficacy is improved by the use of the app and that students' interest and engagement is improved by the use of the app.

The interpretation/use argument makes explicit the many goals of the tablet assessment. Investigating all the assumptions in the interpretation/use argument is a research program that will yield the results needed to assess the validity of intended interpretations and use of the assessment results. Undertaking this program will require a complete set of assessments and a well-planned series of data collection events as well as both qualitative and quantitative analysis. Some of the assumptions, however, should be investigated early in the process to allow the results to inform an iterative development process. As part of our project, we conducted such investigations, whose results allow for an initial look at some of the assumptions in the interpretation/use argument.

## **An Initial Look at Assumptions**

With the prototype app developed, four types of assumptions in the interpretation/use argument could be investigated. First, the domain analysis and domain modeling stages of ECD were intended to serve as backing for the two assumptions underlying the domain definition inference. Second, two assumptions underlying the evaluation inference required data from usability testing to serve as initial backing. Third, two of the assumptions underlying the effectiveness inference required data from the usability study, and two assumptions required data from the focus groups that invited teachers' comments. In the following sections, we report on the usability study and the teacher focus groups, and in the subsequent section the results from these studies are used as initial backing for some of the assumptions.

### ***Usability Study***

The usability study was conducted to obtain some initial data that would reveal whether or not the assessment would function as planned to gather relevant samples of test takers' performance while encouraging them to engage in the tasks. These two characteristics of the assessment will ultimately be critical for making a validity argument for test score interpretation and use. Specifically, making the evaluation inference requires support for assumptions that (a) test takers' performance

is gathered under appropriate conditions to obtain good samples and (b) design features intended to promote learning do not distort the quality of the samples. We therefore gathered data to investigate two questions about the test takers' engagement with the tasks with respect to the evaluation inference:

- 1 Are students able to understand and navigate the tasks as intended?
- 2 What type of evidence does the task provide about (a) learner developmental proficiency and (b) linguistic resources and/or abilities?

In addition, the effectiveness inference in the validity argument will ultimately rest upon assumptions about how appealing and motivating learners find the app. Warrant 3 of the effectiveness inference requires that backing be found for the assumptions that (c) learners perceive the app as a gamelike and aesthetically pleasing environment, and that (e) learners are motivated by the avatars and animated agents. We therefore posed a third question:

- 3 What perceptions do students have about the prototype assessment?

These three questions guided decisions about the types of data to focus on in the usability study, which was the first time that the tablet assessment was trialed by the targeted users.

### **Participants**

Twenty ELLs in a suburban middle school in New Jersey participated in the usability study. Eleven of them were in sixth grade and nine were in seventh grade. Their ages ranged from 11 to 14 years old, and the median age was 12. The participants were diverse in terms of their country of origin: India (8), Dominican Republic (3), United States (3), Egypt (3), Colombia (1), Pakistan (1), and Puerto Rico (1). They also spoke different languages at home. Some of the students reported using only one language at home: Spanish (4), Gujarati (3), Arabic (2), Mandarin (1), Hindi (1), and Marathi (1), and others reported using more than one language at home: Gujarati and English (4), Spanish and English (2), Arabic and English (1), and Urdu and English (1). Most of the students who participated in this study had used an iPad before (18), primarily at school (13), or at home (5). Students reported using the iPad for different purposes: playing games (14), browsing the internet (15), and doing homework assignments (17). Students who had used an iPad typically used it a few times a week (7) or once a week (6); only a few used it on a daily basis (3) or frequently during the week (1). Teachers described the participants' English proficiency levels as either intermediate or high, with approximately half the students at each level; only one student was perceived to be at a low level of English proficiency.

### **Measures**

The measures used in the study included (a) a background questionnaire that teachers completed to provide information about the learners, (b) the prototype app installed on an iPad, (c) a cognitive interview protocol for researchers to complete, and (d) a written survey for learners to complete. The background questionnaire completed by teachers included each learner's grade, gender, age, home language, length of time in the United States, scores on state ELP assessments, and teacher ratings of English language skills. The prototype app, described at length above, consisted of an introductory sequence, two main level tasks, and two diagnostic tasks. The cognitive interview protocol (see Padilla & Benitez, 2014) included observational questions for the researcher to complete while the learner engaged with the app. Observational questions typically related to the learner's ability to follow directions within the app, work independently, use tablet affordances appropriately, acknowledge and utilize feedback, and engage positively with the app. The learner survey included 17 selected-response and three open-ended questions about how much learners liked various aspects of the prototype, their level of motivation, their level of interest in the tasks, and the ease of use of the prototype app. Responses to the first two types of questions were interpreted as indicating the level of backing for assumptions underlying the effectiveness inference, and responses to the second two types of questions were interpreted as backing pertaining to the evaluation inference.

### **Procedures**

Teachers completed the background questionnaires independently, and learners completed the assessment tasks and background questionnaire with the researchers. Each learner completed the prototype tasks individually on an iPad while



**Table 6** Number of Learners at Each of Three Levels of Independence on Each of the Tasks as Indicated by Observers

Ability to work independently	Avatar	Task 1	Task 2	Diagnostic 1	Diagnostic 2
Unable to complete task	0	0	0	1	0
Completed task with some help	2	9	0	10	16
Completed task independently	18	11	20	9	4

supervised by a researcher using the cognitive interview protocol. The researcher also provided assistance to learners as needed to help them understand what they were expected to do. The interviews were audio recorded and were designed to elicit detailed information about usability, difficulty, accessibility, and overall perceptions of the tasks. The majority of the interviews were conducted in English (18) and a few in Spanish (2), depending on a participant's stated preference. At the beginning of each interview, learners were asked to self-report information about gender, age, home language, length of stay in the United States, and prior experience using an iPad. As learners progressed through the app, researchers coded responses to observational questions and took notes. Responses to observational questions (e.g., "Did the student understand what he or she was supposed to do based on the directions/demo?") were coded using predetermined categories (e.g., *yes, partially, no*). After learners completed the prototype, they completed the written survey independently.

The responses to questions in the cognitive interviews were coded to produce quantitative data. These data, as well as data from the teacher survey and learner survey, were summarized using descriptive statistics. Results from these analyses were compiled and summarized across three categories that corresponded to our research questions: (a) learner understanding and navigation of the app (i.e., ability to work independently, follow directions, and utilize tablet affordances), (b) learners' performance on the tasks, and (c) learners' perceptions of the app assessment. They were interpreted with the help of the observational notes that the researchers took as they watched the learners work through the prototype task.

## Results

Results indicated an overall good level of learner satisfaction with the tasks as well as a strong interest in them and motivation for completing them. The results therefore hold promise for finding the support needed for the evaluation and effectiveness inferences in the future. At the same time, results revealed some of the problems that learners encountered in working on the app. These findings point to areas in need of improvement before taking the next steps in development. These results are presented as they address the three research questions.

### *Students' Understanding and Navigation of the Tablet Assessment*

The usability of the app assessment was investigated by evaluating learners' ability to follow directions within the app, their level of independence working within the app, and their ability to engage and interact using tablet affordances. Based on researchers' observations, all the learners were able to work independently to complete avatar selection and Task 2 (see Table 6). Researchers observed that participants had more difficulty independently completing Task 1, Diagnostic 1, and Diagnostic 2. For these tasks, many participants needed some guidance at the beginning of the task, but were able to work independently afterward.

The researchers observed that learners had little difficulty understanding the directions for the avatar selection and Task 2, but had more difficulty understanding the directions for Task 1, Diagnostic 1, and Diagnostic 2 (see Table 7). When asked if it was clear what they had to do in each activity in the learner survey, most learners (18 of 20) either *agreed* or *strongly agreed* that the directions were clear. Thus, although most learners indicated that directions were clear, it was apparent in the observational data that learners did not clearly understand what they needed to do in some tasks, at least initially. The learners may have therefore interpreted "the directions" as including any help they received during task performance.

Based on the coding and notes taken by researchers as part of the cognitive interview, the most frequent point where learners needed help was at the beginning of Task 1 and Diagnostic Tasks 1 and 2. In Task 1, all learners clearly understood that they had to sort source texts into one of two categories. But nine of them misunderstood the intended meaning of the two categories *helpful* and *not helpful*. All nine learners thought the *helpful* category meant the activity depicted in a

**Table 7** Number of Learners Who Understood the Directions as Indicated by Observers

Clarity of directions to students	Avatar	Task 1	Task 2	Diagnostic 1	Diagnostic 2
Did not understand the directions	0	4	0	2	7
Partially understood the directions	1	9	1	7	8
Clearly understood the directions	19	7	19	11	5

source text was helpful or not helpful to the environment. However, what was intended was that learners should decide whether the information in the source text might be helpful in answering their research problem, that is, the case they were supposed to be solving as presented in the app's introductory narrative.

In Diagnostic Task 1, 11 of the learners also had difficulty at the beginning of the task. Observational notes suggested that although learners seemed to know that they had to match words, they did not clearly understand the intended relationship between the words they had to match (e.g., synonym, antonym, related word). In Diagnostic Task 2, 15 learners did not clearly understand the directions at the beginning. Observational notes suggested that learners knew they had to manipulate the words presented in the task, but did not realize that they had to be manipulated to form a sentence. Although some learners needed support at the beginning of each task, they all became more independent as they progressed through the tasks. Despite problems with the digital interface that were apparent in the observational data, most learners (19 of 20) either *agreed* or *strongly agreed* in the participant survey that it was easy to use the iPad.

Thus, no major technology issues were identified in the usability study. There were minor issues with some of the functionality: for example, the scrub feature in Diagnostic Task 1 and the swiping in Diagnostic Task 2. As indicated in the observation notes, in Task 1 two learners encountered a bug in the sample item: The *back* button was not working. This forced the learners to restart the app. In Diagnostic Task 1, the main issue was with the scrub feature. The majority of the learners were not able to figure out how to scrub the wall on their own; the functionality of this feature was not intuitive for them. Also problematic was that it was not clear to the learners where to scrub the wall. In Diagnostic Task 2, a few learners had difficulty swiping or moving the phrases. These learners needed several attempts to swipe the phrases into place. Researchers' observational coding of the learners' technology issues is summarized in Table 8.

#### *Evidence of Students' Developmental Proficiency and Linguistic Resources*

Learners' performance appeared to reveal the intended abilities based on observations of their engagement with the tasks and their responses to items. Descriptive statistics of learners' performance on Task 1, Task 2, and Diagnostic 2 are shown in Table 9, and performance has been summarized as percent of items correct for each round of attempts for each task. Due to the relatively small sample size, classical item analysis was not used to characterize item difficulty and facility. The descriptive statistics show that learners performed well on these tasks, although they performed slightly better in Task 2. In researchers' observational notes, learners reported that they had already seen all the sources in Task 1, so that made it easier for them in Task 2. All the learners answered two of the sources correctly on their first attempt: Video Reporter School and Wind Turbines, indicating these were very easy items. The most difficult source was River Bridge. Only eight learners answered this item correctly on the first attempt, and three more answered it correctly on the second attempt. In Diagnostic Task 2, only two learners answered all the items correctly on the first attempt, but three more learners answered all the items correctly after the second attempt. Most learners had difficulty with two of the sentences: Only eight learners completed "The teacher writes an email" correctly after the first attempt, but only 11 learners completed "The girls listen to the teacher" correctly (eight after the first attempt, three more after the second attempt). This analysis provides limited

**Table 8** Number of Students Who Had Problems Using the Tablet as Indicated by Observers

Technology issues	Task 1	Task 2	Diagnostic 1	Diagnostic 2
No technology issues	18	19	15	14
Some technology issues	2	1	5	6
Many technology issues	0	0	0	0

Note.  $n = 20$ .

**Table 9** Descriptive Statistics for Students' Performance on Tasks 1 and 2 and Diagnostic 2

Task	Variable	<i>n</i>	Mean	<i>SD</i>	Min	Max	Range
Task 1 (300 points)	Percent correct after first attempt	20	77.8	12.99	53.3	96.7	43
	Number of second attempts	20	6.7	3.89	1	14	13
	Percent correct on second attempt	20	77.2	22.53	33.3	100	67
	Percent correct total	20	94.2	7.16	70	100	30
Task 2 (200 points)	Percent correct after first attempt	20	86.3	12.44	55	100	45
	Number of second attempts	20	2.8	2.48	0	9	9
	Percent correct on second attempt	18	78.1	24.41	33.3	100	67
	Percent correct total	20	97.0	3.40	90	100	10
Diagnostic 2 (200 points)	Percent correct after first attempt	20	84.8	13.90	50	100	50
	Number of second attempts	20	3.1	2.78	0	10	10
	Percent correct on second attempt	18	40.2	34.7	0	100	100
	Percent correct total	20	89.3	11.38	60	100	40

**Table 10** Descriptive Statistics for Students' Performance on Diagnostic Task 1

Variable	Mean	<i>SD</i>	Min	Max	Range
Number of words encountered	35.70	16.13	21	84	63
Percent correct after first attempt	72.7	15.17	42.9	96.4	54
Number of second attempts	8.9	4.5	1	17	16
Percent correct in second attempt	51.3	22.95	0	88.9	89
Percent correct total	87.2	7.81	75	98.8	24

Note. *n* = 20.

information about item difficulty and facility, but helped identify several items that were extremely easy or difficult for this sample of learners.

Learners' performance on Diagnostic 1 is presented separately because not all learners completed the same number of items (see Table 10). The number of items attempted by learners ranged between 21 and 84. Overall, learners performed well on this task, although observational notes indicated that many learners commented that some of the vocabulary words targeted by items in this task were difficult. Some of the words learners had difficulty with include *barrel* (18 learners attempted the item; eight answered correctly after the first attempt and four more after the second attempt), *creek* (14 learners attempted the item; four answered correctly after the first attempt and three more after the second attempt), *dessert* (18 learners attempted; nine answered correctly after the first attempt and two more after the second attempt), *gallop* (14 learners attempted; five answered correctly after the first attempt and two more after the second attempt), and *jaw* (17 learners attempted; six answered correctly after the first attempt and three more after the second attempt). One of the items, *vacation*, was extremely easy for this sample of learners. Sixteen learners encountered this word, and all of them answered it correctly on the first attempt.

Across all of the tasks, observational notes indicated that a few learners kept track of the points they were earning in each task. Table 11 summarizes information about the points that learners earned in each task. A total of 300 points could be earned in Task 1 and 200 points in Task 2 and Diagnostic 2. The number of points possible in Diagnostic 1 varied, as not all learners completed the same number of items; in fact, possible points varied between 165 and 790 for learners in this task. In the observational notes, a researcher noted that one learner commented that getting points encouraged him to continue working harder, which was an intended effect of this feedback.

These descriptive statistics of performance, complemented by observational notes and the results of the learner survey, provided some initial evidence pertaining to the evaluation inference, which requires support for the assumption that the scores and feedback from the app are accurate. Accuracy can be achieved only if good samples of learners' performance are gathered through the learners' use of the app. The assumptions about performance being gathered under appropriate conditions and specifically that the design features of the app did not distort the performance of interest also require backing. Such backing was gathered by asking the learners to respond to statements about their level of interest in working on the app and about the task clarity and ease of use. Two of the four statements intended to indicate level of interest were positively worded about the tasks being interesting and about learning occurring. Over 90% of learners *agreed* or *strongly*

**Table 11** Descriptive Statistics for Points Earned

Task	Mean	SD	Min	Max	Range
Task 1	257.50	28.21	180	295	115
Task 2	183.25	14.53	150	200	50
Diagnostic 1	293.00	162.18	165	790	625
Diagnostic 2	175.50	23.45	110	200	90
Total	909.25	188.26	685	1,470	785

Note.  $n = 20$ .

*agreed* with these statements. The two negatively worded statements asked learners to agree or disagree that they got tired of completing the activities and that they got bored. In both cases, over 80% of the learners *disagreed* or *strongly disagreed* with the statements.

### *Students' Perceptions of the Tablet Assessment*

Demonstrating learners' satisfaction and motivation is needed to support Warrant 3 of the effectiveness inference because two of the assumptions are that the users perceive the app as a gamelike, aesthetically pleasing environment and that they are motivated by the avatars and animated agents.

The observational coding shown in Table 12 and notes indicated that most of the learners seemed to be engaged while completing the tasks. Only four of the learners seemed just partially engaged while creating the avatar, as evidenced by their slow movement and frequent looking away from the tablet. Two of the four learners did not explore the avatar personalization options at all and simply left the default options. The rest of the learners seemed engaged while creating the avatar. Some of them seemed enthusiastic and even checked different options before making final selections. Interestingly, two learners chose avatars that did not match their gender.

Observational notes and coding found that most learners seemed very engaged during Task 1. They seemed focused and were paying close attention to what they were asked to do. Some learners read or listened carefully to the sources before responding. Six of the learners did not seem very engaged in Task 1; they were frequently looking away from the tablet and moved slowly. Learners also appeared engaged in Task 2. As learners were already familiar with the sources, they were able to complete this task quickly. Two of the learners were judged to look a little bored. In Diagnostic Task 1, some learners felt the task was too long and got tired of scrubbing the wall toward the end of the task. Some learners also lost interest because they did not know the meaning of some of the words. In Diagnostic Task 2, learners seemed very engaged. Only a few of them seemed frustrated at the beginning, but they became more engaged after they figured out what they needed to do.

Responses to the learner survey corroborated observational coding and notes. The learner survey included 10 selected-response items focused on learners' satisfaction with the app assessment and their motivation toward working with the tasks. Results of learners' responses to these items are shown in Table 13 and were interpreted as indicating the degree of backing for assumptions underlying the effectiveness inference.

Six of the statements made reference to liking some aspect of the task or liking doing the task. The statements were all positively worded, so responses in the *strongly agree* and *agree* categories can be interpreted as providing some initial backing for the assumptions about prospective users seeing the app as pleasing and motivating. For all these statements, at least 90% of the respondents indicated agreement or strong agreement. Only one learner indicated disagreement or strong disagreement with statements about liking (a) how the activities looked, (b) the comic book design, and (c) being a detective. In addition, two learners indicated disagreement with the statement about liking doing activities such as these.

**Table 12** Number of Learners at Each of Three Levels of Perceived Engagement While Completing the Tasks as Indicated by Observers

Level of engagement	Avatar	Task 1	Task 2	Diagnostic 1	Diagnostic 2
Not engaged	0	0	1	0	0
Partially engaged	4	6	2	4	2
Very engaged	16	14	17	16	18

Note.  $n = 20$ .

**Table 13** Learners' Perceptions About the Application and the Tasks

Statement grouped by assessment target	<i>n</i>	Number of respondents for each category			
		Strongly agree	Agree	Disagree	Strongly disagree
<b>Satisfaction</b>					
I liked completing the activities on an iPad.	20	16	4	0	0
I liked how the activities looked on the iPad.	20	12	7	0	1
I always like doing activities like the ones I did today.	20	16	2	2	0
I liked the comic book design of the activities.	18	13	4	0	1
I liked being a detective and solving the case.	18	13	4	1	0
I liked that the activities were part of a story.	18	15	3	0	0
<b>Motivation</b>					
I liked getting stars/points while completing the activities.	20	14	6	0	0
I wanted to complete all the activities.	20	18	2	0	0
I tried my best to complete each activity.	20	17	3	0	0
I would like to do more activities like the ones I did today.	18	16	2	0	0

**Table 14** Number of Learners at Each of Three Levels of Reactions to Feedback as Indicated by Observers

Use of feedback	Task 1	Task 2	Diagnostic 1	Diagnostic 2
Did not react to the feedback	3	5	4	4
Sometimes reacted to the feedback	8	7	5	4
Reacted to the feedback often	9	8	11	12

Note. *n* = 20.

Overall, these responses indicating disagreement constituted a very small percentage of the responses obtained about liking the app, and therefore these results are treated overall as indicating that the app has a very positive potential for obtaining backing for the statement about the app creating a pleasing environment.

The statements intended to gather data about the learners' motivation for working on the assessment yielded similarly positive results. The four positively worded statements obtained responses of *strongly agree* or *agree* from all the learners. Of these agreements, at least 70% indicated strong agreement. These results indicate that it will likely be possible to support the assumption about the motivating nature of the assessment design going forward. Together with the results from the satisfaction statements, the positive motivation responses indicate that future research should be able to provide support for the warrant that learners use the app and explore its features.

Another important aspect of learners' perceptions of the app assessment was related to the feedback it provided. Observational coding indicated that the way learners reacted to the feedback that was provided by the prototype app was similar across tasks (see Table 14). For example, some learners reacted positively to the feedback (e.g., stars, points, sounds) every time they answered an item correctly by becoming happy and even cheering. When an item was answered incorrectly, some learners gasped or seemed surprised; others even put their hands on their face and seemed sad. Observational notes also suggested that a few participants wanted to change their responses, whereas others seemed to try harder immediately after answering an item incorrectly (although many learners did not answer correctly on their second attempt). One learner liked that the app allowed her to correct her mistakes. Another learner tapped on the sources in Task 1 again after getting them wrong, asked the researcher to explain why the response was wrong, and wanted to know what to do to get the correct answer. Researchers also observed that a few learners did not react to the feedback at all, making no attempt to correct their responses or go back to read or listen to the sources again.

Overall, these results indicate promise for supporting assumptions about the quality of the conditions for gathering learners' language samples. The evaluation inference requires support for assumptions that learners' performance is gathered under appropriate conditions to obtain good samples and that design features intended to promote learning do not distort the quality of the samples. The observations of learners working on the tasks indicated that with some additional guidance to get learners started with the tasks, it will be likely that a revised version of the app will be successful in gathering relevant performance. The performance data obtained under conditions where learners had access to such guidance showed that they could perform the tasks well, and their responses appeared to result from their language abilities rather than from other factors. The effectiveness inference in the validity argument requires support for assumptions about how



appealing and motivating learners find the app. The largely positive responses in the perception survey provide initial backing for the assumptions that learners perceive the app as an aesthetically pleasing environment and that learners are motivated by the avatars and animated agents. Although learners had positive perceptions of aspects of the environment that were intended to be gamelike, it is not necessarily clear that learners perceived the app as a game and thus that assumption should be examined by a future study.

### **Teacher Focus Groups**

The primary goal of the teacher focus groups was to obtain feedback from teachers on the prototype assessment tool while it was still under development. This initial feedback was important for identifying any revisions that would be needed to eventually obtain support for two of the inferences in the interpretation/use argument. The interpretation/use argument for the assessment will ultimately require support for assumptions about teachers' interpretation and use of the diagnostic feedback and even about improvement in their teaching. The assumption about improvement of teaching underlies a warrant of the learning inference; that is, that students' learning improves with respect to the ELP Standards. We did not seek evidence for this inference because the prototype was not used in real instruction on an ongoing basis.

Other assumptions about teacher interpretation and use of the diagnostic feedback underlie the effectiveness inference. One of the warrants underlying the effectiveness inference is that teachers target instruction to the KSAs that students need to achieve learning goals. The first assumption is that teachers interpret diagnostic feedback as relevant to developmental standards. The second is that teachers use the diagnostic feedback to plan and modify instruction. All the assumptions about the teachers' interpretation and use of the assessment results as well as their effect on teaching will need to be supported through research on the assessment in use, but at this stage, the goal was to elicit opinions of teachers that would provide an indication of the likelihood of ultimately being able to support such assumptions. We therefore sought data to address these four issues: the teachers' use of the ELP Standards in their classroom, their perceptions of the overall design features of the app and proposed LOA framework, the prototype of the app and currently developed tasks/features, and the information provided by the app for both students and teachers.

### **Methodology**

The data needed to address these questions would have to reflect the perceptions of teachers who were prospective users of the assessment. Therefore, after three groups of such teachers were identified (see next section for a description), a structured focus group methodology was used to present them with a demonstration and hands-on session with the app as well as an opportunity to respond to questions about their perceptions. Three teacher focus group sessions were held in August 2015, with each session lasting approximately 3 hours. Two of these focus groups were conducted on the ETS campus in New Jersey, and the remaining one took place in a school in Iowa. All focus groups used the same structured process for data collection, even though they were facilitated by different members of the research team.

### **Participants**

A total of 23 teachers participated in the study, in groups of 11 (Iowa), seven (New Jersey), and five (New Jersey). Certain eligible participants were recruited by the researchers from local schools. To be eligible to participate in the study, teachers either had to be current middle school English as a second language (ESL) teachers or have extensive recent experience teaching ESL. The participants represented a range of teaching experience, from 1 to over 29 years, and the majority of the teachers held master's degrees. The participants taught students covering a range of English language proficiencies and backgrounds, with some only teaching long-term ELLs and others teaching a combination of long-term ELLs and ELLs with less than 1 year of experience in the United States. Nearly half of the participants reported using some form of electronic device (i.e., a desktop computer, laptop/notebook computer, or tablet) on a daily basis as part of their classroom instruction or organization, with a laptop or notebook being the most popular device. Results for the use of electronic devices as part of formal or informal assessment were more mixed, with only three participants using a desktop or tablet for daily assessment, and four participants using a laptop or notebook. Slightly more participants indicated using these devices for weekly assessment in the classroom, with a total of eight participants using a desktop or laptop

weekly and one participant using a tablet. Ten participants reported using devices a few times a term (two participants with a desktop, five with a laptop or notebook, and three using a tablet). However, many participants noted never using certain electronic devices for assessment (five participants for a desktop and tablet respectively, and three for a laptop or notebook).

### *Research Instruments*

*Premeeting questionnaire.* All participants were asked to complete a premeeting questionnaire about their teaching experience and background, current student populations (i.e., current number of classes and students, average ELP of ELLs, percentage of long-term ELLs and students with less than 1 year of schooling in the United States, and students' native language profiles), as well as their technology use in the classroom for both organizational and assessment purposes.

*Interview protocol.* An interview protocol was used by the researchers to guide the discussion and standardize the focus groups across sessions. The interview protocol was broken up into four sections covering the primary areas of interest for the focus groups. These included discussion questions on the teachers' use of various English language proficiency (ELP) or English language development (ELD) standards in the classroom, an overview of the design framework, a presentation of the app prototype, and the feedback provided by the app for both students and teachers.

*Postmeeting questionnaire.* Teachers were presented with a survey at the end of the focus group session that focused on their individual perceptions of the app prototype and its implementation in the classroom. The survey was divided into four general categories: the app's activities, how they would envision using the app, the usefulness of the app for different purposes, and the potential use of feedback features to help inform instructional decisions. Within these categories, there were individual statements designed to elicit the teachers' perceptions of the app. Participants were asked to agree or disagree with each statement using a Likert scale (i.e., *strongly agree, agree, disagree, strongly disagree, do not know*). Space was also provided for further or more general comments.

### *Procedure*

Participants arrived at the site of the focus groups and were first given the premeeting questionnaire to complete on their own. The session facilitator then welcomed the participants and explained the purpose and goal of the session. The focus groups began with a discussion to elicit how teachers use ELP or ELD standards in their curriculum and instruction. The teachers were presented with an overall description of the assessment design and a demonstration of the prototype and assessment tasks that had been developed thus far, as well as possible prototype tasks for the more advanced levels of the app. Participants were encouraged to share their perceptions about the app, tasks, and enhancement features that had been embedded in the app. This discussion was then followed up by a two-part presentation to provide information on feedback and reporting options for students and then teachers. Participants were asked for their perceptions of the feedback provided by the app and the types of feedback that would be useful for both students and teachers to receive, and the moderator made an attempt to elicit a response from each teacher for each question. Finally, teachers were given the end-of-meeting survey that aimed to elicit further, more specific individualized feedback on the assessment design, app, and its potential use in the classroom. The focus group sessions were recorded and later transcribed by an external transcription service company for further analysis. In addition, research staff took notes during the sessions on the most salient discussion points.

### *Analysis*

The results of the teacher focus groups were analyzed both qualitatively and quantitatively. Data from the background questionnaire and postmeeting survey were analyzed quantitatively using frequency counts and percentages to identify response patterns. Researchers' notes from the focus group interview sessions were used in conjunction with transcriptions of the sessions to identify examples and significant comments of interest relevant to the research questions. The results of the teacher focus groups were analyzed according to the discussion areas in the interview protocol. Therefore, the comments and feedback provided by the teachers are presented on the following points: the teachers' use of ELP or ELD standards in their classroom, their perceptions of the overall design features of the app and proposed learning-oriented assessment framework, the prototype of the app and currently developed tasks and features, and the information

provided by the app for both students and teachers. Finally, any potential issues or concerns are discussed, as well as general suggestions and considerations for further development.

## **Results and Discussion**

The results provided a considerable amount of useful information from a critical set of stakeholders, including many recommendations for revision and the development of extended materials. Here we report on the data pertaining to the evidence suggesting that the teachers would be able to interpret diagnostic feedback as relevant to developmental standards and that teachers would be able to use the diagnostic feedback to plan and modify instruction.

### *Use of English Language Development and Proficiency Standards*

One of the assumptions underlying the effectiveness inference requires that teachers recognize the connection between the feedback and the standards. Therefore, one of the key areas of interest for the teacher focus groups was to ascertain the extent to which the standards play a role in the curriculum and instruction as well as the extent to which the standards impact instructional decision-making. The ELP Standards (Council of Chief State School Officers, 2014), which were used as a reference framework for the app design, were in the process of being implemented for the teachers in Iowa; however, the New Jersey participants were using a different set of standards, the World-Class Instructional Design and Assessment (WIDA) standards. Therefore, all discussion pertaining to the standards for the New Jersey teachers centered on their general use of standards in the classroom, irrespective of the specific standards currently in use.

In terms of the general role of standards in curriculum and instruction, the participants reported that standards were useful overall for placement and to determine a student's level, particularly at the beginning of the year. The Iowa teachers reported primarily using the CCSS to inform their instruction and measure student progress, rather than the English Language Proficiency Assessment for the 21st Century (ELPA21) Standards, whereas the New Jersey teachers noted that their curricula are based on both the CCSS and the WIDA Standards. Some Iowa teachers stated that they only needed to "loosely" know a student's development level. Others indicated that this information was useful for informing content teachers of the proficiency level and capabilities of their ELLs. However, the Iowa teachers observed that there was no specific system in place to determine a student's developmental level and they were still familiarizing themselves with the standards at the time of the focus groups. The New Jersey teachers cited using the official WIDA test and ACCESS scores to determine their ELLs' developmental levels, but added that it would be very useful to know their students' levels in relation to the ELP Standards. In general, the teachers cited a number of different educational programs and diagnostic resources to determine their students' proficiency levels throughout the year, which varied considerably according to the different school districts. Some of these programs provided diagnostic information and were adaptive for ELLs' different proficiency levels, but they were focused on specific skills (e.g., Achieve 3,000 and the Scholastic Reading Inventory, which focused on reading comprehension, and Imagine Learning, which uses the students' native language to test literacy and contains an assessment component). In order to monitor ELLs' reading levels, the Iowa teachers cited using the Basic Reading Inventory and modifying core assessments primarily using textbook assessments from their reading series, *Inside*.

Teachers varied in their reported use of the standards for instruction. Iowa teachers described using the CCSS for learners instead of ELL-specific standards, although the ELP Standards were to be implemented in the upcoming school year. The teachers reported creating their own language objectives based on how the students were progressing according to the CCSS. For instance, one teacher explained, "I kind of just basically go through. .. the regular Common Core standards as well. I just piggyback on that and with what the subject area is, and go through that curriculum as well." The WIDA Standards, and more specifically the WIDA – ACCESS Placement Test, are used for placement at the beginning of the year in New Jersey and are then embedded in the curriculum. Teachers also reported using the standards at the end of the year when they fill out a form with "can do" descriptors to help describe students' abilities, which could then be passed on to high school teachers so they know what to expect of incoming students. These can do statements are based on formal (WIDA) and informal assessments and observations. Teachers also described using informal assessments, observation, end-of-unit assessments, and digital literacy programs (e.g., Northstar) throughout the year to monitor progress, although this varied across participants.

Overall, there were many challenges mentioned by teachers to successful implementation of the standards in their instruction. Generally, teachers indicated that their knowledge regarding the ELP Standards was still a few years behind that of the content teachers. One of the Iowa teachers explained,

We have been trying to marry the standards-referenced grading with no standards. So the expectation is that we do explicitly know where students are at on each standard, but we, as the teachers, are still learning what those standards are.

Iowa participants also suggested that grading proficiency according to the ELP Standards is inconsistent and confusing, with different developmental levels often poorly defined and open to interpretation. Measurements of developmental proficiency are therefore subjective and inconsistent across schools and even across teachers within a school. New Jersey participants also reported the challenges faced by ELLs in their content classes, as mainstream teachers tend to have too many students and are unable to sufficiently focus on the ELLs and their individual needs. Timing was another critical factor in the implementation of the standards. Teachers claimed that they do not have enough time with their students, so they must focus on what the learners need to understand for their content area classes; thus, they are unable to focus on the ELP Standards in the classroom because they need to follow the CCSS in their content classrooms.

Another difficulty for teachers was the fact that classes tend to include students with varying levels of English language proficiency. For instance, some participants observed that the CCSS are more appropriate for advanced level ELLs, but could be challenging for beginner proficiency ELLs. One teacher explained the problems with trying to apply the standards to instruction for beginner level ELLs. She stated,

You just focus on reading, writing, listening, speaking. You just focus on those four skills and always building them and trying to figure out where the kids are and get underneath that and pull them up. And it's as much an instinct role as it is scientific for me.

The teachers also voiced concerns that curricula and assessments designed for beginning level ELLs are often childish and inappropriate for older students, yet many of their beginning level ELLs are older. Teachers expressed a desire for beginner level material without the beginner level children's design. They also explained that it was more challenging to teach English literacy skills to students with low levels of literacy in their native language. As explained by the participants,

The students come with very low literacy levels in their own language and now we have to teach them a new language when they haven't even mastered their language . . . It is very hard for a child if they're not strong in their native language to transfer into another language . . . they don't know the letters in Spanish and now you have to teach them letters in English, so that's the hardest.

Overall, the ELP21 Standards had not had an impact on the teachers' understanding of students' developmental levels. Their lack of familiarity with the ELP21 Standards forecasts a challenge in supporting the assumption that teachers will connect the assessment results with the ELP21 Standards. It remains to be seen how these standards will be adopted in the future, but the nature of the challenges that the teachers raised for their consistent work with any ESL standards needs to be noted in view of the assumption underlying effectiveness of the tablet assessment that teachers will be able to interpret assessment results in view of standards.

### *Prospective Use of Diagnostic Feedback to Plan and Modify Instruction*

Using visual mock-ups, possible feedback options for both students and teachers were presented to participants. Teachers' comments and survey data indicated their interest in having plentiful feedback and, in fact, the more information the app could provide, the more useful it would be perceived by educators. To begin with, teachers were shown a rough sketch of possible student feedback options and asked to provide their general impressions of the design and types of feedback that should be available to students. For instance, the teachers recognized the utility of students being able to review the items they answered incorrectly after completing the level. They also noted that the use of instant feedback is essential, because teachers do not often have the time to provide that type of support for students. In addition, they provided a number of novel ideas about the delivery of feedback.

**Table 15** Teachers' Perceptions on the Type of Feedback the App Should Provide

The app should provide . . .	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
. . . percent correct for each activity.	8	11	1	1	2
. . . a student's overall level with respect to ELP Standards (1–5).	14	6	2	1	0
. . . number of points each student earned in each activity.	6	14	2	0	1
. . . each student's scores on each individual item or question in an activity (right/wrong).	14	8	0	0	1
. . . Aggregated scores across all the activities in a level for each student.	12	9	1	0	1
. . . Scores for different language skills (e.g., listening, reading, vocabulary, grammar)	16	7	0	0	0
. . . Scores for different microskills (e.g., academic vocabulary, comprehension of sentences in present simple tense).	15	7	1	0	0
. . . Lists of words/sentences that students missed in the tasks. <sup>a</sup>	11	10	0	0	1
. . . Feedback specific to the standards (or levels within the standards) that my curriculum is aligned to.	16	6	1	0	0

Note.  $n = 23$ .

<sup>a</sup>One participant did not respond to this item.

In terms of the potential teacher feedback options, participants valued the fine-grained breakdown proposed in the presentation and felt such feedback would be particularly useful to inform future instruction and save valuable time for teachers. Teachers also liked that they could easily identify students' developmental levels, which would be useful to break down classes into smaller groups to focus instruction on targeted skills or pair students at different levels to help each other. The simplicity of the design and use of traffic light colors (i.e., red, yellow, green) were also praised for their efficiency and effectiveness. They also had many suggestions for the teacher feedback.

Feedback elicited from the teachers regarding the information that should be provided by the app is summarized in Table 15. Teachers strongly indicated that the more information and feedback the better; that is, teachers found value and utility in the more comprehensive and detailed feedback on student and class performance that the app could provide. Notably, for every piece of suggested information that the app could offer, more than 80% of responses from teachers were in agreement that the information would be useful. For example, 19 out of 23 teachers believed the app should provide the percent correct for each activity, 20 participants would like to see the number of points earned on each activity, and 22 teachers want student's scores on every individual item in each activity. All the participants would like to see the students' scores for different language skills, and 22 would like to see scores for microskills. Twenty participants would like to see the students' overall level with respect to the ELP Standards, and 22 would like feedback that is specific to the standards to which their curriculum is aligned. Overall, these results indicate that the teachers found the feedback very useful.

The teachers' comments were also solicited on the uses that they saw for the app. Many of the participants agreed that the app addressed a gap in the current market, with no other assessment tools available that drive instruction in such a

**Table 16** Frequencies of Teachers' Responses for Each Level of Agreement Indicating Their Perceptions on the Uses of the App

This app would be useful to . . .	Strongly agree	Agree	Disagree	Strongly disagree	Don't know
. . . place students into different proficiency levels.	18	3	1	0	1
. . . diagnose students' strengths and weaknesses.	22	1	0	0	0
. . . determine if students are meeting the standards.	17	6	0	0	0
. . . place students into instructional groups for differentiated instruction.	20	3	0	0	0
. . . assign a grade or grades.	3	10	4	1	5
. . . formulate teaching plans for individual students.	15	8	0	0	0
. . . identify classroom trends.	15	8	0	0	0
. . . assign as homework.	8	6	3	2	4

Note.  $n = 23$ .



targeted manner. Some teachers stated they would use the app for testing in the classroom, for instance for benchmark assessments. One teacher explained,

Because currently, for example, we have benchmarks in place in our school district and in the department and . . . it takes six to eight weeks to complete a unit. . .and then what we do is an analysis to see what weaknesses and what strengths the kids as a class had. We can look at them individually but we can also look at it collectively. And I think if I use something like this and I do it as a benchmark assessment I can do it . . .each marking period.

The teachers noted in general that the app provided a strong way to integrate diagnostics into classroom work and that this app was the first real diagnostic tool they had encountered. Some teachers claimed they would find this very useful in the classroom, for example, to differentiate instruction for individuals or groups of students. One teacher remarked,

But in order to inform my instruction, something that measures the progress on a regular basis and gives me specific information that is on specific skills will be very beneficial because then I can reform my instruction. I can work on what I want to teach each group and differentiate the different groups according to the needs. It will be great to have something like that at the end that can give me a method of placement.

However, others cited a lack of sufficient time in classrooms as one reason for preferring to use this app as homework or as a supplemental activity. For instance, one teacher explained,

My class sessions are only 40 minutes for each group, so I don't see myself using it as an activity in the classroom. I see using it as a support material, maybe something that's an assignment for the students to do at home.

According to the survey data and discussion, teachers clearly indicated that the app prototype would be useful overall, both as a tool for assessment and to inform instruction. As shown in Table 16, 21 of 23 teachers agreed that the tool would be useful to place students into different proficiency levels. All 23 of the participants agreed it would be useful to diagnose students' strengths and weaknesses and determine if students are meeting the standards. All the teachers would use the app to place students in groups for differentiated instruction and to formulate individual teaching plans for students and identify classroom trends. In fact, the results were more varied in only two categories: using the app to assign grades (with 13 agreeing and the remaining 10 disagreeing or unsure) and assigning the app as homework (with 14 in agreement, five disagreeing, and four unsure).

The participants also discussed areas of caution and consideration when refining the app prototype. To begin with, the fact that students will have varying levels of comfort and proficiency with technology must be considered, particularly when developing instructions for tasks and rubrics to measure constructed (typed) responses. Second, they suggested that not all students would necessarily be familiar with the concept of "being green," and advised that app content should not be "designed exclusively around middle-class values." They also noted that although it is ideal to have written tasks incorporated at the higher levels, grading these responses would be a consideration for teachers in using the app. Although they would be willing to grade students' work with a rubric if provided, they also noted the insufficient time available for extensive marking and expressed a strong interest in incorporating automated scoring technologies.

In addition to these potential concerns with the current prototype design, teachers also provided general feedback and suggestions that will be highly valuable for future task development. First, teachers strongly advocated for the importance of testing certain key skills, namely comprehension, literacy, (academic) vocabulary, and pronunciation (at the lower levels). For example, one participant clarified as follows:

Comprehension skills I think are very important because a lot of [students] come in and they're expected to sort of . . . read at a certain level or be in a certain level and they're just not, not even in their own native language.

Teachers encouraged the incorporation of real-life skills whenever possible within the app, as well as more opportunities for academic and content vocabulary development. Other comments centered on further developing the gamelike enhancement features of the app to increase engagement. Some suggestions included giving the students the opportunity to earn new games or avatar options, the inclusion of more gamelike tasks, and incorporating an adaptive timing

element (i.e., as student performance increases, the timing to complete tasks becomes faster). In general, the diagnostic capabilities of the app were highly praised, and it was even suggested that the app may be useful in helping to distinguish between students who suffer from a language barrier and those with a learning problem that may require special education.

Overall, the teachers' comments indicated promise for finding backing for the assumption that the teachers use the diagnostic feedback to plan and modify instruction. Their enthusiastic engagement with the questions about feedback indicated that they could see clear use for it for both themselves and their students. Still, the limited and varied familiarity and use of standards by the teachers suggests that assumptions underlying the effectiveness inference may require refinement. In particular, the degree to which teachers are able to target their instruction to the relevant KSAs may depend on the language required for performance in particular topical content.

## Summary

The interpretation/use argument presented above expresses the intended interpretations, uses, and consequences of the learning-oriented tablet-based assessment. The warrants and assumptions stated in the argument framework create links from the theoretical frameworks and decisions made during development to the empirical data generated during test use that can potentially serve in support of the validity argument. The first round of such empirical data collection was undertaken as a usability study and teacher focus groups. The results from these small-scale studies provide a basis for optimism about supporting some of the assumptions underlying the evaluation and effectiveness inferences. With respect to evaluation, students' engagement with the app provided good samples of their performance once the students understood what they were supposed to be doing. As for effectiveness, students' responses to the survey as well as their behavior indicated their satisfaction with the aesthetics and gamelike nature of the app. Teachers' enthusiasm about the detailed feedback was very supportive of the assumption that they would use such feedback to plan and modify instruction if it were available to them.

Results also provided some guidance for revisions to the tablet assessment and its interpretation/use argument. Observations of the students as they worked on the app revealed specific areas of confusion for students as they began to work on some of the tasks. These results suggest the need for additional guidance in the form of examples or simulations to show students how the tasks work before they are expected to perform. The teachers' various levels of knowledge and use of the ELP Standards that formed the basis for the assessment raise questions about the feasibility of resting claims about effectiveness on an assumption that the teachers can interpret results with respect to these standards. The suggested areas for revision of the assessment might be undertaken in a future project. What remains for this report is an initial evaluation of the overall test design process.

## Evaluation of the Test Design Process

This project was undertaken as foundational research intended to gain experience with formative language assessment using mobile technologies in middle school content learning. At this stage in the test design process, it is possible to make a preliminary evaluation of the extent to which the first steps of the ECD process were successful in laying the groundwork needed for creating knowledge about tablet-based language assessment for middle school learners in a way that could ultimately serve in the interpretation/use argument. Such an evaluation addresses the question of the value of ECD for capturing the logic of the argument for the assessment design that will be useful as grounds for the validity argument for interpretations and uses of the assessment. This section examines the precise contributions of the ECD process to the validity argument and then summarizes the more general contribution of the process to the design of language tests using mobile technologies.

## Evidence-Centered Design and Validation

The primary motivation for conceptualizing the test development process within an ECD framework is to allow test designers to conceive their assessment work as a method of constructing an argument through their development of design documents, which include rationales emanating from engagement with theory, research, and practice. The purpose of ECD is to develop a test task design that serves as an argument; thus, the success of the test design process

must be evaluated in part by the degree to which it supplies the content required for an interpretation/use argument for the validity of the assessment. In particular, having developed the prototype assessment through an ECD process, we should be able to judge how the processes and products of the ECD framework contribute toward the validity argument and how they help identify what remains to be done to support the validity argument. Table 17 presents an analysis showing how our process and products from ECD will serve as rationales for particular warrants and assumptions of the interpretation/use argument. For each of the inferences and warrants (left column), the ECD contributions appear in either the second or the third columns as providing support (column 2) or some data (column 3). In the fourth column, the assumptions requiring investigation beyond the ECD process are noted. Overall, this table summarizes the specific contributions of the ECD process to the developing validity argument.

The Tablet Project addressed the assumptions underlying the domain definition inference, whose warrant is that observations of performance on the assessment reveal relevant knowledge, skills, processes, and strategies representative of those required for app-based tasks in the middle school classroom. The assumptions for that warrant are that the outcomes from the domain analysis and domain modeling phases of the ECD work are sufficient to inform the subsequent steps in the process. The record of the domain analysis (Chapelle *et al.*, 2015), which is summarized in the second and third sections of this report, serves as backing for the first assumption of the domain definition. The assumption states that domain analysis revealed adequate knowledge about communication and learning in middle school content classes to create useful task patterns, construct definition, and theory of learning as well as to uncover valued conceptions of learning progressions. Having worked through the prototyping, we can now judge that the domain analysis revealed adequate knowledge about communication and learning in middle school content classes to create useful task patterns, construct definition, and theory of learning as well as to uncover knowledge about learning progressions. Because the backing for the assumption is contained in a document, it can be evaluated by others wishing to judge the extent to which it serves as adequate backing.

The task modeling process similarly resulted in a document that serves as support for the assumption that the task patterns developed in the task modeling phase of the ECD process are adequate to serve as a basis for developing tasks reflective of app-based tasks in the middle school classroom. We consider this assumption to be partially supported by this document, based on the fact that the scope of the Tablet Project included development of only one prototype for one level of one of the standards. The task patterns from that document are adequate as a starting point for describing the tasks for the one prototype. However, a more complete judgment of their utility awaits development of tasks for additional standards and tests.

The ECD process allowed us to begin to investigate the plausibility of assumptions underlying the evaluation inference, which was based on the warrant that scores and feedback provide teachers and students with accurate information. The usability testing conducted as part of assessment implementation provided initial but limited evidence pertaining to the first and second assumption: that performance is gathered under appropriate conditions to obtain good samples and that design features intended to promote learning do not distort the quality of the samples. In fact, results of the usability study also proved informative for proposing recommendations that would improve the quality of the performance samples by giving more guidance to students about how to respond to tasks. Similarly, during usability testing, the scoring rules (*i.e.*, percent correct for each task) were assessed for their success in providing relevant scores; feedback was carefully examined for its appropriateness relative to the scores and for its accuracy. These initial explorations of scoring and feedback provide a starting point for constructing tasks that can stand up to future investigation, even though we do not consider the results from this exploration as backing because it was not carried out and recorded systematically. These explorations of the tasks during usability testing therefore provide a first step in developing backing for the other three assumptions underlying the evaluation inference: that scoring rules in the evidence model are designed and implemented to provide relevant scores (CAF development), that feedback reflects the outcomes of the scoring rules (assessment implementation), and that feedback is perceived as accurate to experts. Such research would need to evaluate the accuracy and appropriateness of scores and feedback for students using the app and to gather the judgments of experts about the scoring and feedback.

The ECD process laid the groundwork needed to provide backing for one of the assumptions underlying the generalization inference in the domain modeling phase by creating the task designs. One of the warrants for generalization is that tasks are designed according to consistent specifications. The domain modeling results shown in the third section of this report provide the basis for test specifications, and therefore provide some backing for this assumption, even though

**Table 17** Inferences in the Interpretation/Use Argument for the Tablet Assessment With the Assumptions That Have Been Partially Investigated During Evidence-Centered Design

Inferences — Warrants	Sufficient backing yielded from the ECD process (ECD phase)	Some relevant data obtained during the ECD process	Assumptions not yet addressed
<p>Domain definition — Warrant: Observations of performance on the assessment reveal relevant knowledge, skills, processes, and strategies representative of those required for app-based tasks in the middle school classroom.</p>	<p>Assumption 1: Domain analysis revealed adequate knowledge about communication and learning in middle school content classes to create useful task patterns, construct definition, and theory of learning as well as to uncover valued conceptions of learning progressions (domain analysis).</p>	<p>Assumption 2: Task patterns are adequate to serve as a basis for developing tasks reflective app-based tasks in the middle school classroom (domain modeling).</p>	<p>N/A</p>
<p>Evaluation — Warrant: Scores and feedback provide teachers and students with accurate information.</p>		<ol style="list-style-type: none"> <li>1. Performance is gathered under appropriate conditions to obtain good samples (assessment implementation: usability testing).</li> <li>2. Design features intended to promote learning do not distort the quality of the samples (assessment implementation: usability testing).</li> <li>3. Scoring rules in the evidence model are designed and implemented to provide relevant scores (CAF development).</li> <li>4. Feedback reflects the outcomes of the scoring rules (assessment implementation).</li> <li>5. Feedback is perceived as accurate to experts (assessment implementation).</li> </ol>	<p>NA</p>
<p>Generalization — Warrant 1: Scores are consistent across tasks, forms and occasions. Warrant 2: Feedback is consistent across multiple instances of the same error.</p>			<p>Warrant 1: Assumptions 2, 3 Warrant 2: Assumptions 1, 2</p>
<p>Explanation — Warrant: Scores and diagnostic results reflect intended aspects of the construct of academic language performance.</p>			<p>Assumptions 1, 2, and 3</p>
<p>Extrapolation — Warrant: Scores and diagnostic results are relevant to performance required for success in middle school content area courses.</p>			<p>Assumptions 1, 2, and 3</p>
<p>Effectiveness — Warrant 1: Teachers target instruction to KSAs that students need to achieve learning goals. Warrant 2: Students focus study on abilities they need to gain mastery of the standards. Warrant 3: Students use the app to explore its features.</p>			<p>Warrant 1: Assumptions 1, 2 Warrant 2: Assumptions 1, 2, and 4 Warrant 3: Assumptions 1, 2, and 4</p>
<p>Learning inference — Warrant 1: Students' learning improves with respect to the ELP Standards. Warrant 2: Students become more autonomous learners.</p>		<p>(For Warrant 3) 3. Learners perceive the app as gamelike and aesthetically pleasing environment. (For Warrant 3) 5. Learners are motivated by the avatars and animated agents.</p>	<p>Assumptions: all</p>

they would need to be used to create more tasks to be more thoroughly evaluated. The other assumptions underlying this warrant (Assumptions 2 and 3) require empirical research using a complete form of a test. The second warrant for the generalization inference is that feedback is consistent across multiple instances of the same error. The assumptions are about having sufficient opportunity to have frequent error feedback and the consistency of the error feedback. Backing for both of the assumptions will require analysis of the performance of the feedback from a large number of learners over a sufficient period of assessment use.

Some of the data gathered during the usability testing might be considered relevant to the explanation inference, whose warrant was that scores and diagnostic results reflect intended aspects of the construct of academic language performance. However, more research results are needed to serve as adequate backing for these inferences. The third assumption about the relationship between scores on the tablet assessment and those on other tests of the same construct would require a correlational study designed to assess that question.

Similarly, the findings from the teacher focus groups seem promising for eventually supporting the extrapolation inference. The warrant for that inference is that scores and diagnostic results are relevant to performance required for success in middle school content area courses. The teacher focus groups provide some initial positive data because of their positive opinions about the content and tasks in the assessments. Overall, the teachers saw the diagnostic results from the assessment as relevant to linguistic resources required for tasks in the content classroom. Although the data from these focus groups were overall supportive of the fit between the assessment and classroom instruction, support for the three assumptions would need to come from research systematically investigating these aspects of the assessment. Such research would need to analyze correspondences between the assessment tasks and those in the classroom from the perspectives of teachers and students who were using the assessments in their classes.

The effectiveness inference received some initial support from the usability testing results. The third warrant for the effectiveness inference is that students use the app to explore its features. Aspects of two of the assumptions were clearly supported by the data: that the learners perceive the app as an aesthetically pleasing environment and that learners are motivated by the avatars and animated agents. The findings from one usability study are not adequate backing; these assumptions would need to be investigated again with students using the assessment for their real classroom learning.

The other assumptions underlying the effectiveness inference are in need of additional research. A particularly thorny issue for the effectiveness inference was revealed by the teacher focus groups. Warrant 1 states that the teachers will target their instruction to the KSAs that the students need to achieve learning goals. The first assumption for that warrant is that the teachers interpret the diagnostic results as relevant to developmental standards. In the focus groups, the finding was that the standards used to develop the assessment were unfamiliar to the large majority of teachers, who were just beginning to work with the CCSS (but not the ELP21 Standards) or who worked with a different set of standards altogether. If any effectiveness claim to be made about the assessment really relies on a standards-related interpretation of assessment results, this assumption will require additional attention by ensuring that the teachers using the assessment are also knowledgeable about, and working toward, achievement of standards.

The ECD process did not provide any data for the learning inference, which would need to be investigated in a classroom context where the assessment materials are used regularly over time because the warrant is that students' learning improves with respect to the ELP Standards. The assumptions encompass two hypothesized antecedents to students' learning: Teaching improves by becoming more efficient, systematic, and responsive to students' needs; and learner autonomy increases. These assumptions would need to be investigated in a classroom using both quantitative and qualitative data.

## **Knowledge About Tablet-Based Assessment**

By working through the first stages of an ECD process, the Tablet Project was intended to yield knowledge that would be useful for designing tests delivered through tablets. In order to frame this broad inquiry, we posed four questions, which are revisited here to evaluate the extent to which the questions were answered through the project. The first three questions were addressed in the domain analysis, and the last one was addressed by going beyond domain analysis to domain modeling, development of the CAF, and assessment implementation.

In response to the question of how mobile technologies are actually being used in middle school classrooms in the United States in 2013, we found a considerable variety of practices. The use of tablets varies by school, classroom,



teacher, day, and lesson to the extent that identifying typical tablet use for communication and learning was not possible. Based on the existing and prospective practices of tablet use in the classroom, we focused our investigation on identifying the innovative practices that took advantage of the affordances of the technology. To do so, we reviewed the scholarship on language learning and technology as well as educational technology for middle school. We also used our contacts to identify classes where the tablets would be used in a manner that would be informative for our project. Based on our interpretation of both the published work that looks ahead by reporting on exploratory use of tablets primarily in research projects and on observations of selected classrooms engaged in tablet-based learning, we were able to identify current and likely future practices. These results are included in the domain analysis document.

The second question was how the strategies and abilities called for by tablet-based communication and learning could be characterized. We addressed this question by developing a theoretical framework that stated our understanding of the abilities underlying performance on tablet-based tasks. We used the framework as a basis for defining the specific constructs intended to explain score meaning. Our observations suggest that learners' interactions with tablet-based tasks draw upon abilities beyond those used for other forms of communication and interaction. Our reviews of the literature on computer-mediated communication and learning supported this observation. To develop a theoretical framework that would capture the strategies and abilities students use to engage in tablet-based learning, we began with the framework that was being used in the other programs for ELLs at ETS. We expanded that framework to include not only linguistic meaning-making resources but also nonlinguistic meaning making to take into account that students' interactions with tablets often take place through tapping on the screen and other nonlinguistic moves. We also added technologies that are used in the four contexts of language use in the schools. The framework for multimodal communication was created to address this question.

The third question was how particular features of design used in educational apps can serve for learning and assessment purposes. To address this question, the multimodal construct framework helped in conceptualizing task features relevant to assessment of particular abilities. For learning, we included in the literature review LOA and formative assessment in addition to selected laboratory-based work on learning through technology. The result was the theory of action framework, in which the action includes observable improvements in learning and teaching, and the theory includes the features of design that are expected to prompt such improvements through specific hypothesized action mechanisms.

The fourth question was how tablet-based formative assessment tasks can be designed to take into account the constructs to be measured and the intended learning. This question calls for concrete plans for developing tasks, an assessment framework describing the underlying models, and the implementation of an actual prototype task. We accomplished these steps in the design by following the guidance of ECD, and the results are reported in the task modeling and implementation documents. More specifically, we found that the project required a complex process of pooling knowledge in assessment, language, learning, and technology. Mislavy (2011) advised that "collaboration and interaction from the beginning of the design process is needed" (p. 21). The collaboration should include

... users (who understand the purposes for which the assessment is intended), domain experts (who know about the nature of the knowledge and skills, the situations in which they are used, and what examinees do to provide evidence), psychometricians (who know about the range of situations in which they can model data and examine and compare its evidentiary value), and software designers (who build the infrastructure to bring the assessment to life). (p. 21)

Our experience confirms this observation.

## Summary

The goal of ECD is to systematize the process of assessment design in a way that produces the knowledge required to make design decisions as well as a trace of the rationales that connect that knowledge to the actual test design. In the Tablet Project, ECD served well in both these capacities. The knowledge is presented in the domain analysis, task modeling, and task implementation documents, which include rationales for the decisions that we summarized

throughout this report. The rationales are only as good as they are useful as a basis for the interpretation/use argument. As shown above, the rationales developed through ECD are able to serve in the interpretation/use argument to help make explicit the inferences, warrants, and assumptions entailed in assessment use. Moreover, the learning-oriented focus of the third question led to the development of some task features that were not intended to have a substantial impact on the construct of measurement, but nevertheless have intermediate- and long-term intended effects on learning. By specifying these components and their effects in the theory of action and then including them in task design, we were then able to specify “effectiveness” and “learning” inferences in our interpretation/use argument. Specification of such an interpretation/use argument is the first step in moving toward the research required for validation.

## Conclusion

The Tablet Project was undertaken as foundational research to create a prototype assessment for ELLs in middle school, and as such, it was intended to increase understanding of the nature of the challenge itself. We began with a number of broad questions about current and projected use of mobile technologies in middle school classrooms, the abilities and strategies required for their use, the features needed in an app used for both assessment and learning, and the concrete design features for making an app that would succeed with users in this context. Following current practice, we approached the project through the use of ECD. Therefore, the project also offers an example of the use of the ECD process for test design. Demonstrating the use of ECD for a variety of purposes is important because ECD is a general framework for designing tests across subject areas and contexts, and as such, the guidance it provides is limited for any individual project. Moreover, as we have shown, examples of ECD use also require an evaluation of the success of the approach in view of how it meets its goal of providing a basis for the interpretation/use argument for the assessment. Working within the general ECD framework, we were able to shed light on four issues: the use of the mandate to place parameters on the ECD process, creation of a construct framework that takes into account technology, description of a domain expected to exist in the future, and conducting an initial project evaluation by developing an interpretation/use argument for evaluating the prototype.

One issue that arises in the use of ECD is the need for a means of specificity in how to make decisions about the scope and nature of the design activities. In this project, we benefited from several example descriptions of ECD use in other projects, but many issues required project-specific guidance to resolve. ECD alone does not direct test designers about what the domain analysis should include or how many design patterns the domain modeling should specify, for example. An ECD domain analysis directs test designers to survey the relevant knowledge in the domain of interest to test users, but a domain can be defined and surveyed in many different ways. Our solution of specifying and drawing upon the project mandate provided parameters on the ECD process. The test mandate offers direction on how the survey should proceed by specifying the context of the assessment, such as the intended score meaning, who will use test results, and the purpose. For example, in the Tablet Project, the mandate included the need for scores to have meaning in a context where learning was the critical issue. The domain analysis therefore needed to include a developmental perspective that would be relevant to the test score users. The implications of the mandate for the domain analysis will differ depending on its detail and scope, but the general observation that the ECD process requires a specified mandate in order to be usefully engaged is probably relevant for all testing projects.

A second issue that arose in the project was the need to create a construct framework that takes into account the technology-mediated nature of language use. Implicit in the mandate was the assumption that the tablet technology contributed new, different, and consequential affordances to the processes of interest in language assessment and learning. The domain analysis revealed considerable support for this view, particularly in work on multimodal literacies that theorize the integral role of technologies in communication and learning. Such a view is also consistent with perspectives on construct definition in language testing that recognize the role of the contextual factors at play during test performance and therefore need to be taken into account in score interpretation. Given the prevalent view in language testing that context interacts with ability, technology-mediated tasks and assessments could benefit from an explicit recognition of the digital resources that may facilitate or encumber performance. More broadly, one can argue that perspectives from the study of multimodal literacy need to be added to an updated theory of communicative language ability to include the affordances offered and the skills required in contexts where tablets and other mobile technologies play an important role in communication. The question of whether the technology

contributes to construct meaning, and therefore should be considered construct relevant or construct irrelevant for any intended score use, needs to be considered explicitly. Including technologies used in the contexts of interest in our multimodal framework was useful as a first step toward explicitly recognizing them during the test design process.

A third issue that is relevant to many test design projects for technology-mediated tests is that the technologies being used within the domain of interest are likely to be undergoing change. If a construct framework that includes technology is integral to the construct meaning, a change of technologies can be consequential for test design. By including tablet technology, the mandate of this project presented the need for us to address the issue of describing a domain of language use that did not yet exist. The domain analysis could not rely on the observations of random classroom technology use and a survey of academic language ability. Instead, it had to devise a methodology for looking into the future expected normal practices in such classrooms. Our methodology relied on the study of the professional literature in the use of technology for learning in addition to some carefully selected classroom visits where mobile technologies were being used, even though these technologies were not yet widespread.

A fourth important issue we addressed in this project was the need for an evaluation of the test design process. ECD is conceptualized as a process whose outcomes should be plans and prototype tests in addition to being a basis for a future validity argument. To evaluate the success of our project in meeting these goals, we initiated a project evaluation by developing an interpretation/use argument for evaluating the prototype. Such an interpretation/use argument serves as the basis for the validity argument; therefore, the role of the ECD activities in the validity argument should be apparent. In addition, the adequacy of the ECD results is tested by the quality of the interpretation/use argument that is outlined and the degree of support offered for the domain definition inference. In this way, the project should be informative to future development efforts by showing how well the ECD process serves as input to development of the interpretation/use argument. ECD and argument-based validity are regularly presented as compatible, and accordingly, validation studies sometimes cite the ECD process as backing for a domain definition inference. But the purpose of ECD is to reveal the rationales that link knowledge of the domain with task design decisions. Moreover, the purpose of argument-based validity is to make explicit the inferences underlying the interpretation and use of test results by revealing their logical connections and underlying assumptions. In this process, then, the rationales revealed in ECD need to be made explicit in the validity argument. The Tablet Project did this.

Our project tackled these four issues in the use of ECD for test design and, in doing so, presented novel approaches to the challenges test designers meet using ECD for technology-based tests. The changing role of technology creates a challenging agenda for English language assessment, whose purpose is to deliver not only efficient assessments but also relevant ones. This challenge, which prompted the launch of our Tablet Project in 2013, is becoming more and more familiar as tests continue to move online, and therefore this project may prove instructive to others tackling such design projects.

## References

- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice*, 18, 5–25. <https://doi.org/10.1080/0969594X.2010.513678>
- Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, 7, 161–186. [https://doi.org/10.1016/S0959-4752\(96\)00015-1](https://doi.org/10.1016/S0959-4752(96)00015-1)
- Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International*, 44, 57–66. <https://doi.org/10.1080/14703290601081332>
- Chapelle, C., Cho, Y., Hutchison, A., Lee, H.-W., Schmidgall, J., & Wain, J. (2015). *School literacy demands in tablet-based learning: First steps in test design for assessment of communication practices in future middle schools in the United States*. Unpublished manuscript.
- Chapelle, C. A., Enright, M. E., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. London, England: Routledge.
- Coiro, J., Knobel, M., Lankshear, C., & Leu, D. J. (2008). Central issues in new literacies and new literacies research. In J. Coiro, M. Knobel, C. Lankshear, & D. J. Leu (Eds.), *Handbook of research on new literacies* (pp. 1–21). Mahwah, NJ: Lawrence Erlbaum Associates.
- Council of Chief State School Officers. (2014). *English language proficiency (ELP) standards*. Retrieved from [https://ccsso.org/sites/default/files/2017-11/Final%204\\_30%20ELPA21%20Standards%281%29.pdf](https://ccsso.org/sites/default/files/2017-11/Final%204_30%20ELPA21%20Standards%281%29.pdf)

- Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University.
- Educational Testing Service. (2009). *Research rationale for the Keeping Learning on Track program*. Princeton, NJ: Author.
- Halliday, M. A. K. (2004). *An introduction to functional grammar* (3rd rev. ed.). London, England: Hodder Arnold.
- Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford, England: Oxford University Press.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. <https://doi.org/10.1111/jedm.12000>
- Kress, G. (2003). *Literacy in the new media age*. New York, NY: Routledge
- Kress, G. R., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London, England: Arnold.
- Lankshear, C., Knobel, M., & Curran, C. (2012). Conceptualizing and researching “new literacies.” In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 863–850). Hoboken, NJ: Wiley Blackwell. <https://doi.org/10.1002/9781405198431.wbeal0182>
- Leu, D. J., Kinzer, C. K., Coiro, J. L., & Cammack, D. W. (2004). Toward a theory of new literacies emerging from the Internet and other information and communication technologies. In R. B. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1570–1613). Newark, DE: International Reading Association.
- Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365. <https://doi.org/10.1111/j.1467-9922.2010.00561.x>
- Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302. <https://doi.org/10.1017/S0272263109990520>
- Marazano, R. J., & Simms, J. A. (2013). *Vocabulary for the common core*. Bloomington, IN: Marzano Research.
- McNamara, D. S., Jackson, G. T., & Graesser, A. (2010). Intelligent tutoring and games (ItaG). In Y. K. Baek (Ed.), *Gaming for classroom-based learning: Digital role-playing as a motivator of study* (pp. 44–65). Hershey, PA: IGI Global.
- Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment* (CRESST Report 800). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. <https://doi.org/10.1111/j.1745-3992.2006.00075.x>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62. [https://doi.org/10.1207/S15366359MEA0101\\_02](https://doi.org/10.1207/S15366359MEA0101_02)
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199–218. <https://doi.org/10.1080/03075070600572090>
- Padilla, J. L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*, 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*, 544–555. <https://doi.org/10.1037/0022-0663.92.3.544>
- Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33–40. <https://doi.org/10.1037/0022-0663.82.1.33>
- Pintrich, P. R., Roeser, R. W., & De Groot, E. A. M. (1994). Classroom and individual differences in early adolescents' motivation and self-regulated learning. *The Journal of Early Adolescence*, *14*, 139–161. <https://doi.org/10.1177/027243169401400204>
- Popham, W. J. (2008). Formative assessment: Seven stepping-stones to success. *Principal Leadership*, *9*(1), 16–20.
- Ranta, L., Lyster, R., & DeKeyser, R. (2007). A cognitive approach to improving students' oral language abilities: The awareness-practice-feedback sequence. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 141–160). New York, NY: Cambridge University Press. <https://doi.org/10.1017/CBO9780511667275.009>
- Schmidgall, J., Lopez, A., Blood, I., & Wain, J. (2015). *Tablet technology and learning-oriented English language assessment for young learners*. Unpublished manuscript.
- Schunk, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist*, *40*, 85–94. [https://doi.org/10.1207/s15326985ep4002\\_3](https://doi.org/10.1207/s15326985ep4002_3)

- Valezy Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). Amsterdam, The Netherlands: Benjamins. <https://doi.org/10.1075/lllt.13.09val>
- Wiggins, G. (2012). Seven keys to effective feedback. *Educational Leadership*, 70(1), 10–16.
- Wolf, M. K., Everson, P., Lopez, A., Hauck, M., Pooler, E., & Wang, J. (2014). *Building a framework for a next-generation English language proficiency assessment system* (Research Report No. RR-14-34). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12034>
- Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences*, 8, 211–238. [https://doi.org/10.1016/S1041-6080\(96\)90015-1](https://doi.org/10.1016/S1041-6080(96)90015-1)
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41, 64–70. [https://doi.org/10.1207/s15430421tip4102\\_2](https://doi.org/10.1207/s15430421tip4102_2)

### Suggested citation:

Chapelle, C. A., Schmidgall, J., Lopez, A., Blood, I., Wain, J., Cho, Y., Hutchison, A., Lee, H.-W., & Dursun, A. (2018). *Designing a prototype tablet-based learning-oriented assessment for middle school English learners: An evidence-centered design approach* (Research Report No. RR-18-46). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12232>

**Action Editor:** Donald Powers

**Reviewers:** John Norris and Tanner Jackson

ETS, the ETS logo, MEASURING THE POWER OF LEARNING., and TOEFL are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at <http://search.ets.org/researcher/>