International Journal of Engineering Education Vol. 33, No. 1(B), pp. 441–452, 2017
Printed in Great Britain

0949-149X/91 $3.00+0.00

Developing an Engineering Design Process Assessment Using Think-Aloud Interviews*

MELTEM ALEMDAR[1], JEREMY A. LINGLE[1], STEFANIE A. WIND[2] and ROXANNE A. MOORE[1]
[1] The Center for Education Integrating Science, Mathematics, and Computing (CEISMC), College of Sciences, Georgia Institute of Technology, 817 West Peachtree St., Suite 300, Atlanta, GA, 30303, USA.
[2] College of Education, University of Alabama, Box 870231, Tuscaloosa, AL, 35487, USA. E-mail: meltem.alemdar@ceismc.gatech.edu; jeremy.lingle@ceismc.gatech.edu; stefanie.wind@ua.edu; roxanne.moore@gatech.edu

Early exposure to engineering has been found to help students in their decision-making regarding engineering education and career pathways. Subsequently, an NSF-funded project is underway that is focused on development of an engineering curriculum for students in grades six through nine. The Engineering Design Process (EDP) frames this curriculum. The current study presents the validation methods and results of a multiple-choice assessment created to measure students' understanding of the EDP. The utilization of Think Aloud Interviews and the application and analysis of qualitative coding schemes for the purpose of systematically gathering evidence about the psychometric quality of the assessment are described. Findings from this study support the validity of the EDP assessment through evidence of alignment between the intended skills and the skills elicited in the student interviews.

Keywords: engineering design process; assessment; validation; cognitive interviews

1. Introduction

The engineering education community and leaders in the field of technology education have identified the important role K-12 engineering education plays in the success of postsecondary engineering education [1]. Hence, an early exposure to engineering can help students make informed decisions about engineering as a career path. The United States is no exception, where the role of K-12 engineering education continues to be of national interest [2]. Through a National Science Foundation (NSF) funded project, Georgia Institute of Technology partnered with a public school district to bring engineering curriculum to students in grades six through nine. In this project, middle school students explore Science Technology Engineering Mathematics (STEM) Innovation and Design in engineering technology courses.

In order to guide instruction related to engineering design, the curriculum utilizes the Engineering Design Process (EDP) as a sequential and/or iterative process. A variety of EDP models have been used as guiding frameworks for engineering curricula that vary in terms of specific terms, order, and sequences [3] (e.g., see [4]). In order to contribute to the vital conversation surrounding development of valid assessments for engineering education [5], the current study describes efforts to develop a valid and reliable assessment to inform the development and implementation of an engineering curriculum. Evidence Centered Design (ECD) [6] is used as a framework for assessment design. Using a mixed-method approach, quantitative and qualitative techniques are used together to explore student responses to a multiple-choice engineering design assessment as evidence to strengthen the validity argument for the instrument and guide revisions to individual items. The quantitative component focuses on exploring the psychometric characteristics of an engineering design assessment using Item Response Theory.

The qualitative component, which is the major focus for the current study, includes the use of Think-Aloud Interviews (TAIs) to gather evidence about student conceptions of engineering concepts and their rationale for selecting answer choices. TAIs are useful for identifying cognitive processes and knowledge structures in which students engage as they complete a test [7]. Additionally, the qualitative data were used to explore student responses to the assessment items regarding student cognitive processes and perceptions of item difficulty drivers.

This study illustrates the use of TAIs as a systematic method for gathering evidence about the psychometric quality of an EDP assessment. It then presents empirical results from TAIs that indicate the cognitive processes that students employed as they responded to items on an engineering assessment. Lastly, the TAI results are summarized in terms of the ways students defined and utilized engineering design concepts.

The major purpose of this study is to explore student responses to multiple-choice (MC) engineering assessment items in order to gain a more complete understanding of student conceptions of

* Accepted 15 October 2016.

engineering design and to inform revisions to and development of new assessment items. Following the methodology described by Hamilton, Nussbaum, and Snow [8] and DeBoer, Lee, and Husic [9], concurrent think-aloud interviews and retrospective probes were used to gather evidence about student conceptions of engineering and their rationale for selecting answer choices. This study is guided by two major research questions:

1. Do the piloted engineering design process items elicit evidence of the intended cognitive processes?
2. What item features contribute to the perceived difficulty of the piloted engineering design process assessment items?

This study contributes to the field of engineering education in several ways. First, it provides validation information regarding an assessment of the Engineering Design Process among middle school students. Second, the study provides an illustration and guidance toward a rigorous, systematic approach to validating assessment instrument through the use of Think-Aloud Interviews.

2. Theoretical framework

The Committee on Developing Assessments of Science Proficiency in K-12 issued a set of recommendations for the design of assessments aligned with the Next-Generation Science Standards [10] that reflect an emphasis on the integration of practices, crosscutting concepts, and disciplinary core ideas in science education. The committee called for the use of frameworks for assessment design that "provide a methodological and systematic approach to designing assessment tasks" [10, p. 52]. The final recommendations emphasize the role of evidence as a key aspect of assessment design frameworks that is needed in order to "support the validity argument for an assessment's intended interpretive use and to ensure equity and fairness" [10, p. 81]. Following this recommendation, the theoretical framework for this study draws upon principles from Evidence-Centered Design (ECD).

Recognizing that assessment is an evidentiary reasoning process, it is important to use a systematic process while designing an assessment. ECD is a framework for assessment design that focuses on: the role of evidence in developing assessment tasks and contexts that elicit a particular construct, the intended inferences from assessment scores, and the nature of the evidence that supports the inferences [10]. This process starts by defining as accurately as possible particular aspects of a content domain—in other words, the ways in which students are supposed to know and understand the content. Examples from the current study include the use of the EDP as a cognitive model (see Figure 1). Additionally, the claims that one wants to be able to make about student knowledge play a critical role for the purpose of the assessment [11]. This study focuses on the *evidence model* component of ECD, in which empirical evidence is examined to support the interpretation of responses to assessment tasks as indicators of student achievement in terms of a construct [12, 13].

2.1 Using think aloud interviews in evidence-centered assessment design

Think Aloud Interviews (TAIs) are recommended for developing a cognitive model of task performance as a method for gathering validity evidence to support the interpretation and use of an assessment [14]. As described by Leighton [15], developing a cognitive model of task performance is a necessary step because "this model is the type that researchers develop to confirm empirically that students are employing the expected knowledge and skills on the items being developed" [15, p. 8]. Development of a cognitive model was important in this study because the cognitive processes related to the EDP have not been tested or described in the literature. In terms of the evidence model component of ECD, the role of cognitive processes that students employ while completing assessment tasks, and the degree to which these processes reflect the intended construct are very important. Ferrara et al. [16] refer to this type of evidence as *item construct validity evidence*. One approach to gathering validity evidence is through the use of cognitive labs, or think-aloud interviews, during which students either concurrently or retrospectively describe the cognitive processes they employed when responding to assessment tasks [17]. Therefore, TAIs are useful for identifying the cognitive processes and knowledge structure when students perform a task. During the TAIs, students are directed to freely "think aloud" as they respond to an item, which provides researchers information about the cognitive processing performed when responding to the item.

3. Methods

Data for this study were collected using an EDP assessment that includes 18 multiple-choice (MC) items with distractors that were constructed to reflect common student misconceptions about different stages of the EDP. An example of such a misconception in understanding of the EDP is the perception of the design process as linear, rather than an iterative process that requires revisiting prior design decisions and evaluating alternative solutions. Another example of a misconception is

students' ignoring constraints and requirements for a design due to a preferred solution by the individual student [18]. The items were developed based on pre-existing engineering assessment items [19] and subject-area expert review. Further, the items were aligned to one or more stages in a conceptual model of the EDP used in the curriculum; this conceptual model is illustrated and defined in Figure 1. Each stage was measured by at least two items: four stages of the EDP were measured by four items each; one stage, Problem Understanding, was by far the most commonly appearing stage and was expected to be elicited by eight of the assessment items. The instrument was pilot-tested in January 2014, and the post-test and cognitive interviews were conducted in May 2014.

A cognitive interview protocol was adapted from protocols described in previous TAI studies [9]. This procedure uses EDP MC items as stimuli. The interview began with the researcher modeling "thinking aloud" while answering an example item. The student then reads and chooses a response while verbalizing their thinking. Following the "think aloud" portion of the interview, the interviewer asks the student to elaborate on their understanding of the item, strategies and sources of knowledge used to select a correct response, and rationale for eliminating distractors using a semi-structured interview protocol. Hamilton, Nussbaum, and Snow [8] also stated that this type of interview procedure, combined with multiple-choice items, allows researchers to discover student reasoning processes and strategies for responding to MC items, sources of knowledge applied to MC items, and differences in reasoning and strategies between successful and unsuccessful students. The semi-structured interview protocol was pilot-tested for validity purposes. The pilot test assisted the research team in identifying weaknesses within the interview design, and allowed the researchers to make necessary revisions and estimations of time requirements prior to the implementation of the study. The primary changes made to the design as a result of the pilot test was to reduce the length of the introductory statements describing the interview protocol process, as well as to remove several probes from the interview protocol that were determined to be redundant.

3.1 Case selection

The pre-test scores of students enrolled in the engineering classrooms (see [20]) were used to construct a stratified sample for the cognitive interviews. This was accomplished by placing each student into one of three performance-level groups of approximately equal size (low, medium, or high) based on their pre-test achievement estimates. Similarly, items were categorized according to their difficulty (easy, moderate, or difficult) based on student pre-test performance estimates. A total of six item sets were created so that each set included an easy, moderate, and difficult item. Students from each of the performance-level categories were purposefully selected for interviews for each item set, ensuring that students from all achievement levels provided data for items of all difficulty levels.

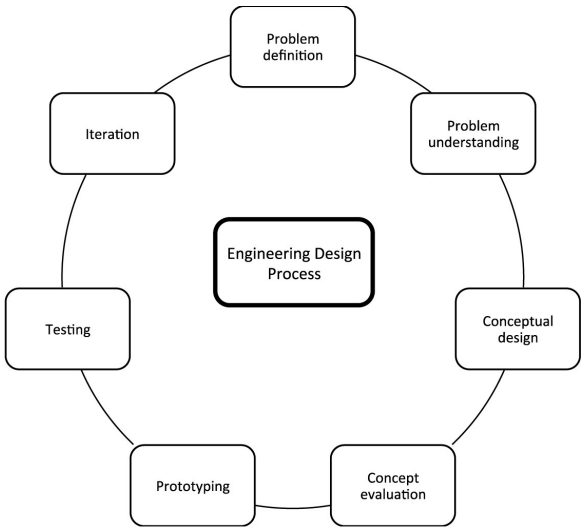
3.2 Participants

Participants in the qualitative component of the EDP assessment development study included a sample of 44 students (four students did not want to participate in the interviews) enrolled in a public middle school with approximately equal numbers of students from each sixth, seventh, and eight grade levels. Prior to the interviews, all students had participated in the semester-long engineering curriculum; therefore, students had a basic understanding of engineering vocabulary to be able to participate in the interviews. Interviews were conducted by a group of eight educational researchers who practiced protocol administration prior to conducting interviews. All interviews were conducted in English and lasted approximately 20 minutes. In order to keep the interviews to a 20-minute time period, only three EDP assessment items were included per interview. All students who participated in the interviews were proficient in English.

3.3 Data analysis

Prior to conducting interviews, a preliminary coding framework was developed based on the framework described by Kaliski et al. [14]. Particularly, codes within four major categories were specified: (Category A) cognitive processing, (Category B) difficulty drivers, (Category C) test-taking behaviors, and (Category D) miscellaneous. This coding framework provides a systematic method for categorizing student responses that aligns with the conceptual model of engineering that is included in the curriculum. Using the framework, four trained researchers independently coded verbal reports using nVivo® software. Codes were modified following initial analyses to better reflect the scope of responses. This study focuses on Category A and Category B.

The cognitive processes that students employed as they responded to the items were captured by the codes in Category A. This category included five codes. First, the code "Factual Recall" was used when students recalled specific facts or definitions to answer a question rather than using a specific cognitive skill (e.g., Student response: "It mostly sounds like the right definition for it."). Second, the code "Engineering Design Process" was used when



Working Definitions for Engineering Design Process Stages	
Problem Definition	The first stage of the design process, during which the engineer /designer identifies a specific problem to be solved. Rather than focusing on possible solutions, the goal during this stage is to clearly identify the need that is to be addressed.
Problem Understanding	During the problem understanding stage, the engineer /designer identifies critical aspects of the problem that will affect their success. Specifically, the engineer/designer should identify: 1) the client/market/customer and their requirements and preferences; 2) the functions that the designed solution must carry out; 3) the constraints or resource limitations that will affect the solution; and 4) the engineering requirements for the problem that directly affect customer needs (e.g. material strength).
Conceptual Design	During the conceptual design stage, the engineer/designer identifies possible solutions to the problem at a conceptual level, without focusing on the technical details. This stage is also referred to as "ideation" or "brainstorming."
Concept Evaluation	Using the requirements, the engineer/designer identifies which design(s) are the most likely to meet the customer's needs before proceeding with more detailed designs/ models/ prototypes/ final products. Designs are sometimes evaluated using a matrix in which the probability of success with respect to each customer requirement is rated for each design.
Prototyping	During the prototyping stage, the engineer/designer develops a testable version of a designed solution.
Testing	During the testing stage, the engineer/designer performs multiple tests on the prototype(s) to see if the design met the requirements and constraints. Statistics are often used to monitor and describe testing results, when appropriate (e.g., a 90% success rate for launching into the specified area).
Iteration	During iteration, the engineer/designer draws upon past designs to inform future designs. Iteration is not a clearly defined or separate stage; rather, iteration characterizes the engineering design process in that various stages may be frequently updated and revisited when new knowledge about the problem or proposed solutions is acquired.

Fig. 1. Conceptual Model (Subset of Coding Category A).

a student's response indicated consideration of one or more of the EDP stages. The EDP stages are defined operationally in Figure 1. If the student made reference to the EDP or if engineering reasoning was demonstrated without clear indications of the stage, the code "Evidence of Intended Skills" was used. Examples of these statements made by the participants included:

> You could go through the design process in your head and think about what your final design must include and then that has to be your goal. That would be the answer choice if you're using the engineering.
>
> Because we were designing problems to design solutions to fix problems.
>
> Because you're going to research through the process, but that's not really your goal—to research.
>
> Because you need to know that to actually fix the problem. You can't just go to the end and think about problems similar to it.
>
> Because engineers have to figure something out, like figure out how to improve stuff.

Third, the "Guessing" code was used when the students indicated that they did not have sufficient knowledge to determine the answer. Fourth, stu-

Table 1. Summary of Coding Analyses

Item Number*	Interview Count (N)	Intended Skills/ EDP stage(s)	Evidence of Alignment with Intended Skills/EDP Stages (*N*)	Evidence of Alignment with Other Skills/EDP Stages (*N*)	Difficulty Drivers Identified (*N*)
1	8	General EDP	*N* = 6	Conceptual design (*N* = 2); Iteration (*N* = 1); Problem definition (*N* = 1); Prototyping (*N* = 1); Testing (*N* = 1)	*N* = 0
2	7	General EDP Problem understanding Problem definition	*N* = 4 *N* = 0 *N* = 0		Item vocabulary (*N* = 3); Quality of distractors (*N* = 4)
5	6	Concept evaluation Prototyping	*N* = 1 *N* = 1	General EDP (*N* = 5)	Item vocabulary (*N* = 3); Quality of distractors (*N* = 5)
6	8	Problem definition Problem understanding	*N* = 0 *N* = 0	General EDP (*N* = 7)	Quality of distractors (*N* = 1)
7	7	Problem understanding	*N* = 0	General EDP (*N* = 3)	Quality of distractors (*N* = 7)
8	9	Problem definition Problem understanding	*N* = 4 *N* = 7	Concept evaluation (*N* = 8); General EDP (*N* = 3)	Degree of familiarity (*N* = 3); Item vocabulary (*N* = 1); Quality of distractors (*N* = 4); Stimulus material (*N* = 1)
10	9	Problem understanding	*N* = 4	Concept evaluation (*N* = 2); Iteration (*N* = 1); Testing (*N* = 1); General EDP (*N* = 3)	Degree of familiarity (*N* = 1); Item vocabulary (*N* = 3); Quality of distractors (*N* = 1)
11	7	Problem understanding	*N* = 2	Concept evaluation (*N* = 1); General EDP (*N* = 4)	Item vocabulary (*N* = 3)
13	8	Concept evaluation Conceptual design Iteration	*N* = 4 *N* = 0 *N* = 0	Problem understanding (*N* = 1); Prototyping (*N* = 1); Testing (*N* = 1); General EDP (*N* = 3)	Item vocabulary (*N* = 1); Quality of distractors (*N* = 1); Stimulus material (*N* =)
14	4	Problem understanding	*N* = 0	Conceptual design (*N* = 3); Problem definition (*N* = 1); Prototyping (*N* = 2); Testing (*N* = 1); General EDP (*N* = 3)	Quality of distractors (*N* = 2)
15	5	Testing Iteration	*N* = 1 *N* = 1	General EDP (*N* = 4)	Degree of familiarity (*N* = 1); Item vocabulary (*N* = 1); Length of item (*N* = 1); Quality of distractors (*N* = 3)
17	4	Concept evaluation Testing	*N* = 3 *N* = 0	Problem definition (*N* = 1); Problem understanding (*N* = 1)	Item vocabulary (*N* = 1)
18	6	Concept evaluation Problem understanding Testing	*N* = 4 *N* = 0 *N* = 0	Conceptual design (*N* = 1); General EDP (*N* = 3)	Degree of familiarity (*N* = 1); Item vocabulary (*N* = 2); Quality of distractors (*N* = 3)
20	7	Testing Iteration	*N* = 5 *N* = 1	Problem understanding (*N* = 3); General EDP (*N* = 3)	Degree of familiarity (*N* = 2); Quality of distractors (*N* = 3)
21	7	Problem definition	*N* = 0	Concept evaluation (*N* = 7)	Quality of distractors (*N* = 2)
22	7	Concept Design	*N* = 0	Concept evaluation (*N* = 1); Conceptual design (*N* = 3); Problem definition (*N* = 3); Problem understanding (*N* = 4); General EDP (*N* = 5)	Quality of distractors (*N* = 8)
23	7	Prototyping	*N* = 5	General EDP (*N* = 5)	Item vocabulary (*N* = 2); Quality of distractors (*N* = 5)
24	7	Iteration	*N* = 2	General EDP (*N* = 6)	Quality of distractors (*N* = 6)

* *Note:* Only the 18 items that were administered to the students in the current study are included in the table.

dent responses were coded for "Process of Elimination" if they used a strategy to eliminate answer choices. Fifth, student responses were coded as "Background Characteristic" if students referenced personal experiences or background characteristics when choosing an answer, such as referencing a family member who is an engineer.

The codes in Category B, Difficulty Drivers, describe item features that students indicated as increasing or decreasing the difficulty of an item but were not directly related to the content of the item. "Item Length" was coded because it was hypothesized that longer item stems or answer choices increase the difficulty of an item. Second, comments about additional material included with the item, such as graphics or charts, that made the item more difficult were coded as "Stimulus Material." This code was also used when the stimulus

material served to clarify the item, and thus made it easier. Third, ''Degree of Familiarity'' was used if students indicated a lack of previous exposure to the information, usually scenario-based (e.g., lack of familiarity with airplane travel). This code appeared rarely in the interviews (see Table 1). In instances where a student indicated lack of familiarity, interviewers were trained to probe to determine if this lack of familiarity affected student understanding of the scenario. Fourth, ''Quality of Distractors'' was applied when students stated that multiple response options appeared plausible or if some distractors were easy to eliminate. Fifth, the ''Vocabulary'' code was used when the meaning of a word was not known. Sixth, student statements that indicated a misunderstanding of any part of the item were coded as ''Misunderstanding.'' For frequency of occurrence of each code see Table 1.

Although not examined in this study, coding categories C and D are described. Category C focused upon student test-taking strategies, including ''Process of Elimination'', ''Rereading'' or restating portions of the item, misreading words that impacted choice selection (''Misread''), changing an answer choice after making a selection (''Change Answer''), and using clues within the item to select or eliminate a response option (''Scaffolding within the Item''). Following Kaliski et al. [14], the ''Process of Elimination'' code is included in Category A and Category C because this action can function as both a cognitive process and a test-taking strategy. Category D included ''Miscellaneous'' codes, and included indications of correct or incorrect responses (''Correct Response'' or ''Incorrect Response), any apparent difficulty with the think-aloud process (''Difficulty thinking aloud''), required prompts from the researcher (''Researcher Prompt''), or student comments indicating that the stimulus material (text, graphics, or charts) was irrelevant (''Stimulus Material Irrelevance'').

3.4 Coding process

The coding was completed in three rounds. First, in order to be certain of a common understanding of the codes and to address potential definition refinements, four researchers used the initial framework to code the transcripts related to three assessment items. If evidence of a code appeared at all for an assessment item, the entire interview was coded. Following the first round of coding, the researchers met and refined the initial definitions. As a result, the initial code definitions were expanded to include comments related to both item stems and answer choices (initially the code verbiage was focused upon answer choices), the definition of ''Background Characteristic'' was revised to exclude students' experiences of EDP through their engineering curriculum, and a new code was created to capture references to the EDP that were not clearly indicative of a specific stage (''Evidence of Intended Skills'').

In the second round of coding, the same process of using a code per item if evidence of it was apparent was applied using the refined codes. The purpose of this whole-item coding was to explore the frequency of codes per item in order to identify those items that warranted further exploration. The third and final round involved more traditional qualitative coding in which a code could be used multiple times per item. Coding in this manner provides a more in-depth understanding of the areas of primary interest, namely student cognitive strategies, implicit and explicit use of the EDP, difficulty drivers, and misconceptions in student understanding of the EDP process.

4. Results

Before examining the cognitive interview transcripts in depth, counts of correct and incorrect responses were calculated for each EDP multiple-choice item. Findings indicated that the number of correct responses corresponded to the achievement-level groups that were assigned based on the pre-test performance, with more incorrect responses appearing among ''low performing'' group members, and fewer incorrect responses appearing among ''high performing'' group members.

In this section, results from the qualitative analysis of the cognitive interviews are summarized in terms of the guiding research questions for this study. A discussion of conclusions from these findings follows.

Research Question 1: Do the piloted engineering design process items elicit evidence of the intended cognitive processes?

The degree to which an assessment item elicited EDP knowledge as intended was made apparent through the analysis shown in Table 2. This table presents a summary of results from Round 2 of the qualitative analysis, in which codes appear once per item per interview. The table presents the frequency of transcripts with identified intended skills related to the previously described cognitive processes (including references to the EDP) and the difficulty drivers. This illustration reveals that all assessment items elicited at least one intended skill related to the EDP, and all items except Item 1 were associated with at least one difficulty driver.

To provide an example of the examination of alignment to intended skills, consider Item 5 in Table 1, for which six cognitive interviews were conducted. This item was designed to elicit the *Concept Evaluation* and *Prototyping* stages of the

Table 2. Summary of observations about EDP as a cognitive model and Difficulty Drivers

EDP Stage/Code	Types of Responses	Example Student Response
Problem definition	Explicit reference to problem definition as a method for clarifying the appropriate next steps in an engineering design challenge.	I used defining the problem . . . If you didn't understand the problem up here, then you couldn't really answer this down here because you would be confused.
	Explicit reference to problem definition as a defined step in the EDP that they learned in class.	(I used) defining (the problem) because our teachers usually use the word define it, they tell us what to do or they have a lot of paper and we read the paper and it help us define the problem.
	Implicit reference to problem definition by identifying or focusing on the specific needs of a customer/client.	The reason why I chose C is because your main goal is to be able to allow dogs to have enough air to fly safely for eight hours and be sound proof enough that passengers cannot hear barking dogs. You want to be able to meet these and keep them and solve the ABC Airline problems. That's what C is saying to design a new container that solves the Airlines problems.
Problem understanding	Implicit reference to problem understanding by focusing on identifying and understanding requirements and constraints for an engineering design problem.	You need to review the requirements and restraints (constraints) of the problem you are solving. Like, they want the dogs to have enough air to fly for eight hours, they want to be sound proof, and it needs . . . and the requirements need to be, um, the size and how much it costs, and it can't be poisonous to dogs. That's what it is saying, those are the requirements and problems you are solving.
	Implicit reference to problem understanding by focusing specifically on the requirements and constraints in terms of the customer/ client.	Because you want to see, you want to find soundproof materials so that the customers can be happy on their trip.
	Implicit reference to problem understanding by focusing on the functions that the designed solution must carry out.	It's not all about the cost and materials. Instead it's about what it needs to do and stuff.
	Implicit reference to problem understanding by focusing on engineering requirements that affect customer needs.	Since they are so close together, you need to try and make sure you have soundproof materials that are good to make sure they . . . here because they're so close together.
	Focus on relative ordering of problem understanding within the EDP.	You don't conduct research on things related to the problem (first), you want to think about what the problem is.
	Focus on the importance of problem understanding within the EDP.	If you just make random changes to see if the problem goes away, then you're not really considering the fact that there's a problem at all, because you don't know where the problem is . . . and so if you don't know where it is, then how can you know if you're fixing the problem that's in the game, instead of just . . . you know, making random changes.
Conceptual design	Implicit reference to conceptual design where students described the relative ordering of brainstorming or ideation within the engineering design process.	You can't start building a new game until you brainstorm a game into your head, until you know what it is.
	Description of conceptual design (brainstorming) as an essential process of engineering design that is used for generating ideas.	You know how you first want to build a catapult, but you don't know what the design you want to do is, so first you'd have to brainstorm possible designs.
Concept evaluation	Students indicated use of concept evaluation when they described the importance of specific customer needs or criteria when considering the quality of a solution.	You're not just focusing on soundproof materials because you've got the other things to work on . . . he wants you to build something good, but you don't need to focus on the strongest thing because you need all the things right here, all the requirements.
Prototyping	Students described the use of prototypes as a part of iteration.	You shouldn't build a full-scale. You should do a little mini one and test is out to see if it would work.
	Students described the use of prototypes as a method for understanding potential solutions.	(Create) a prototype or building a simple drawing of it so you could get a simple base idea about what you are going to do without adding all the extras to it yet.
Testing	Students explicitly referenced the concept of testing as an essential method for several aspects of successful engineering design.	He should test it more and see what the problem would be. If he documents it, he will get the answer for why it messes up.
	Students described testing as a method for diagnosing problems with a design in order to inform iteration.	Because if the game stops working at level 3, then that means something isn't going right, so he would have to carefully test it . . . in order to know what's not working, and how to solve the problem, and like when he makes the results . . . when he checks the results then it'll be easier for him to look over them without him getting messed up, or losing where he stopped at.
	Students described testing as a method for comparing potential solutions.	You have to test it to see if it will work, and she has to test her different versions of the device. Of each material.
	Students described testing as a method for verifying a solution.	Just because they say it can clean a hundred carts in thirty-five minutes don't actually mean that it can, so she needs to test it to see.

Table 2. (*continued*)

Difficulty Drivers Code	Type of Responses	Example of Student Response
Iteration	Students noted examples from personal or class experiences with iterating on a design.	It's like when we made a prototype of a cradle design for the catapult. Ours wouldn't throw the ball into the safe zone. So we changed up the design but still kept it the same a little bit and it started working.
	Students referred to iteration as a method for ensuring adherence to design requirements if an original design was unsuccessful.	If you keep your original design and you begin the game and no one makes it, you could end up having a bad game and you wouldn't be able to come back into the carnival.
	Students described the concept of improving upon previous designs, rather than starting over, within the context of a design challenge.	I would keep running it and running it and make changes and see would that help it and if it does I would stick to that instead of trying to do the process over. I would iterate the process I already have and just keep doing it until it works and if it doesn't work at a certain time, then I'll start over.
Length	Student indicates that the length of item stimulus material makes an item difficult.	I almost picked that because it was too many words and it got confusing.
	Student indicates that the length of the item stem makes it difficult.	
	Student indicates that the length of an answer choice makes it difficult.	
Stimulus material (graphics, charts, etc.)	Student indicates that the stimulus for an item (e.g., graphics, charts, etc.) makes the item difficult.	I didn't get out what this little thing is, or what it's supposed to be.
	Student indicates that the length of the stimulus makes the item difficult.	
Degree of familiarity	Students have not had an opportunity to learn the content, making the item less familiar.	I never heard (of this). I don't know how they do the shopping carts. I didn't know they use this much money to do this.
	Students indicate that the item context is unfamiliar.	
Quality of distractors	Student indicates that some distractors were easy to eliminate.	A and D were sort of kind of alike, so they sort of confused me.
	Student indicates that two or more distractors appear to be plausible options.	
Item vocabulary	Student indicates that the item was difficult as a result of vocabulary (in the stimulus, item stem, or answer choices).	I'm not going to say D because I don't really know what that second word is.
Misunderstanding	Student does not understand the item stem.	I was confused with C because I really didn't understand it.
	Student does not understand an answer choice.	

EDP. Examination of the codes indicated that references to *Concept Evaluation* were made in one interview, and references to *Prototyping* were also made in one interview. Five students made general reference to the EDP, although they did not specifically identify a stage. For example, one student responded that: ''An engineer fixes a lot of things and it can be like a bunch of stuff. It can be ways to improve the safety of cars.'' In this statement, the student seemed to be describing the role of engineering as improving products, suggesting at least some understanding of the EDP, but with too little detail to allow identification of a specific stage.

Considering these findings, the researchers examined the specific sections of the relevant transcripts coded during Round 3 and determined that students were inconsistently interpreting the EDP stage presented in the scenario. This discovery resulted in refinements to both the vocabulary used and the scenario for this item.

Research Question 2: What item features contribute to the perceived difficulty of the piloted engineering design process assessment items?

The researchers also examined the difficulty drivers that were coded for each item. Evidence of perceived difficulty drivers were typically identified as part of the semi-structured, retrospective probe portion of the TAI, when the researchers asked students about why an item was easy or difficult. The difficulty drivers identified are listed in Table 1 in the rightmost column. To continue with Item 5 as an example, the item-level coding results revealed that three students found the vocabulary challenging, and five students found the quality of the distractors to be an issue. Upon review of the specific coded sections of the transcripts for this item, the researchers discovered that language inconsistencies between the scenario and the response options caused some confusion (using

"requirements" in the scenario and "criteria" in the response options), which was corrected in subsequent iterations of the item.

5. Conclusion

Findings from this study suggested that Think Aloud Interviews, along with the coding framework based on the EDP, proved to be a valuable method for gathering evidence about the psychometric quality of this EDP assessment. Overall, results suggested that the EDP assessment items were generally eliciting the intended skills. Specifically, findings from the qualitative analyses indicated alignment between the intended and observed EDP skills for the new assessment items examined in this study. This finding provides item construct validity evidence for the EDP assessment. Second, certain EDP skills (e.g. *Concept Design*) were less frequently observed than others (e.g. *Concept Evaluation*). This variation in the eliciting of certain stages of the EDP is hypothesized to be the result of the emphasis that is placed on these stages in the engineering curriculum, which has a greater focus on evaluating specific designs and less focus on initial development of those designs. This curricular emphasis likely impacts the students' ability to identify characteristics of more familiar stages. Third, there were several features that students perceived as contributing to the difficulty of the items, primarily related to difficulty choosing between multiple response options they perceived as correct. Results for several items indicated potential issues related to item clarity and vocabulary; these issues were typically simple refinements to make, once identified. However, some vocabulary issues identified in this study contributed to the formative evaluation of the curriculum. For example, student confusion over certain terms considered primary to the understanding of EDP, such as "constraints" and "iterations," were used for curricular adjustments and related teacher professional development.

Acknowledgements—The authors would like to thank Anna N. Holcomb from CEISMC at Georgia Tech for contributing to the data collection. This material is based upon work supported by the National Science Foundation under Grant No. 1238089. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Georgia Institute of Technology. For additional information about the research related to this project see https://ampitup. gatech.edu

References

1. C. Hailey, T. Erekson, K. Becker and M. Thomas, National center for engineering and technology education, *The Technology Teacher*, **64**(5), 2005, pp. 23–26.
2. R. L. Carr, L. D. Bennett and J. Strobel, Engineering in the K-12 STEM Standards of the 50 U.S.States: An Analysis of Presence and Extent, *Journal of Engineering Education*, **101**(3), 2012, pp. 1–26.
3. C. J. Atman, O. Eris, J. McDonnell, M. E. Cardella and J. L. Borgford-Parnell, Engineering Design Education, in A. Johri and B.M. Olds (eds), *Handbook of Engineering Education Research*, Cambridge University Press: New York, 2014, pp. 201–226.
4. R. Bailey and Z. Szabo, Assessing Engineering Design Process, *International Journal of Engineering Education*, **22**(3), 2005, pp. 508–518.
5. K. A. Douglas and S. Purzer, Validity: Meaning and Relevancy in Assessment for Engienering Education Research, *Journal of Engineering Education*, **104**(2), 2015, pp. 108–118.
6. R. G. Almond, L. S. Steinberg and R. J. Mislevy, Enhancing the design and delivery of assessment systems: A four-process architecture, *Journal of Technology, Learning, and Assessment*, **1**(5), 2002, pp. 3–63.
7. K. Ercikan, R. Arim, D. Law, J. Domene and S. Lacroix, Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews, *Educational Measurement: Issues and Practice*, **29**(2), 2010, pp. 24–35.
8. L. S. Hamilton, E. M. Nussbaum and R. E. Snow, Interview procedures for validating science assessments, *Applied Measurement in Education*, **10**(2), 2009, pp. 181–200.
9. G. DeBoer, H.-S. Lee and F. Husic, Assessing integrated understanding of science, in Y. Kali, M. C. Linn and J. E. Roseman (eds), *Coherent Science Education: Implications for Curriculum, Instruction, and Policy*, Teachers College Press: New York, 2008, pp. 153–182.
10. National Research Council, Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12, in J.W. Pellegrino, M. R. Wilson, J. A. Koenig, A. S. Beatty (eds), National Academics, Washington, DC, 2014.
11. J. W. Pellegrino, Assessment of science learning: Living in interesting times, *Journal of Research in Science Teaching*, **49**(6), 2012, pp. 832–841.
12. R. J. Mislevy and G. D. Haertal, Implications for evidence centered design for educational assessment, *Educational Measurement: Issues and Practice*, **25**, 2006, pp. 6–20.
13. *Large Scale Technical Report: Leveraging Evidence-Centered Design in Large-Scale Test Development,* http://ecd.sri.com/downloads/ECD_TR4_Leveraging_ECD_FL.pdf, Accessed 25 September 2015.
14. P. Kaliski, M. France, K. Huff and A. Thurber, Using Think Aloud Interviews in Evidence-centered Assessment Design for the AP World History Exam, *American Educational Research Association*, New Orleans, IL, April 2011.
15. J. P. Leighton, Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing, *Educational Measurement: Issues and Practice*, **23**(4), 2004, pp. 6–15.
16. S. Ferrara, T. G. Duncan, R. Freed, A. Vélez-Paschke, J. McGivern, S. Mushlin, A. Mattessich, A. Rogers and K. Westphale, Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment, *American Educational Research Association*, San Diego, CA, April 2004.
17. K. A. Ericsson and H. A. Simon, Verbal reports as data, *Psychological Review*, **87**(1), 1980, pp. 215–251.
18. C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg and J. Saleem, Engineering Design Processes: A Comparison of Students and Expert Practitioners, *Journal of Engineering Education*, **96**(4), 2007, pp. 359–379.
19. T. J. Moore and S. S. Guzey, EngrTEAMS engineering content assessment for grades 4–8. Engineering to Transform the Education of Analysis, Measurement, and Science through a Targeted Mathematics-Science Partnership, *National Science Foundation*, Twin Cities, MN, 2013.
20. S. A. Wind, M. Alemdar, J. D. Gale, J. A. Lingle and R. Moore. Developing an engineering design process assessment using mixed methods: An illustration with Rasch measurement theory and cognitive interviews, *American Educational Research Association*, Chicago, IL, April 2015.

Appendix

Interview Protocol

General Notes:

- The interviewer's script is written in italics.
- Parts II and III of the procedure should be repeated for each assessment item.
- You are encouraged to validate the student during the interview: "*This is exactly the information we need; you are doing great*."
- You may follow-up student responses with further probes for student understanding. If you ?recognize that you no longer understand a student's reasoning, clarifying questions should be asked. In addition, questions specific to the topic (e.g. vocabulary questions) or to the distractors may be asked. ?
- Depending on student responses, it may be necessary to skip questions in Part III if they would be redundant.

Introduction: *Thank you for participating in this research activity today. We are developing engineering questions for middle-school students and we need your help to find out if they are good questions. Even though you may not know all the answers, you can still help us figure out if the questions are fair and easy to understand. We are not going to tell your teacher how you do, and you won't get a grade.*

To find this out, I will be conducting a think aloud interview with you. You've been asked to participate because you are a middle school student, and we are going to be doing these interviews with other kids who are about your age.

Now, let me tell you more about what exactly a think aloud interview is. Basically, you will complete the questions out loud so that we know what you are thinking when you are answering the questions.

Let's talk more about what this means. Think back to when you took the engineering assessment. When you took the test, you read each question one at a time, thought about the answer silently to yourself in your mind, and selected your answer. The difference between that and what happens in a think aloud interview is that now you think about this question out loud. You speak any and all thoughts that run through your mind. Today, we will give you some questions, and for each question, we would like you to think aloud, saying what comes to mind, while you determine which option you will choose.

When you took the engineering assessment the first time, your goal was to score as high as possible. Now, the goal is different: the questions are being evaluated, not you. There are no right or wrong thoughts. Say everything that is on your mind. It is important to explain how you reach the answer you select, and any problems you may have along the way to determining your answer.

I. Part One: Concurrent Think-Aloud

Before we begin, let's look at an example. First, I will demonstrate thinking aloud with a question. Then, I will give you some questions and ask you to think aloud while you answer them. When you finish thinking aloud, I am going to ask you some more questions about how you came up with the answer. You can ask me to explain any words or situations that may be unfamiliar or confusing. Do you have anything to ask before we get started?

Now that you understand what is involved, are you still okay participating?

1. Demonstrate the think-aloud process with the following item:
[Select an example item, and compose your think-aloud script to demonstrate thinking aloud.]

2. Present the first item. Then ask the student to complete the think-aloud procedure.

- If the student is having trouble thinking aloud, the researcher should wait for 15 seconds of silence before using neutral probes such as: "*Keep on talking*", or "*Please continue*", or "*Go on*." The researcher should not say probes like "*What are you thinking?*"
- If the student continues to have trouble thinking aloud, ask: *Could you tell me in your own words what the question is asking?*
  - Remind the student to go back and re-read the question if they need to.?

3. As time allows, repeat the think-aloud procedure (step 2) for the remaining items. *?*

II. Part Two: Retrospective Probes

1. *Which answer did you choose? Why did you choose it?*
   a. Probe for words or diagrams to which the student paid particular attention.
   b. Probe for the student's thinking behind the response:
      i. *Did you use an engineering idea to answer the question?*
2. *Were there other answer choices that you almost chose? If so, why?*
   a. If so, probe for test-wiseness:
      i. *What helped you decide not to select that choice? Were there any clues?*
3. *Were there any answer choices that you did not even consider? If so, why?*
   a. Probe for test-wiseness:
      i. *Did you use a strategy to make an educated guess?*
      ii. *What clues helped you know that you could eliminate one/some of the choices?*
4. *Was there an answer choice you were expecting to see, but did not? If so, what was it?*
   a. If so, probe for prior conceptions:
      i. *Why did you expect to see that as an answer choice?*
      ii. *Is one of the answer choices close to what you expected to see?*
5. *Were there any words or diagrams you did not really understand, or situations that made the question confusing? Do you think something would be confusing to your classmates?*
   a. Probe for comprehensibility:
      i. *What is confusing about it?*
      ii. *What does it mean to you?*
   b. If there is a picture, ask:
      i. *Was the picture useful?*
      ii. *What did you use from the picture to help you answer?*
6. *Are you familiar with the situation that is presented in the question?*
   a. Probe for appropriateness of task context:
      i. *Does the situation seem realistic to you?*
      ii. *Is the situation interesting?*
      iii. *Is the situation easy to understand?*
7. *Where did you learn about the topic in this question?*
   a. If the student does not mention their engineering class, ask: *Is there anything that you have done in your engineering class that reminds you of this question?*

Meltem Alemdar is Assistant Director and Senior Research Scientist at Georgia Tech's Center for Education Integrating Science, Mathematics, and Computing (CEISMC). Dr. Alemdar has experience assessing programs that fall under the umbrella of educational evaluation, including K-12 educational curricula, K-12 STEM programs after-school programs, and comprehensive school reform initiatives. Across these evaluations, she has used a variety of evaluation methods, ranging from multi level evaluation plan designed to assess program impact to methods such as program monitoring designed to facilitate program improvement. Dr. Alemdar's research interest includes development and validation of assessment tools within K-12 engineering education, and attributes related to student's intention to persist in STEM education. She received her Ph.D. in Research, Measurement and Statistics from the Department of Education Policy at Georgia State University.

Jeremy Lingle is currently a Research Faculty member at Georgia Tech's Center for Education Integrating Science, Mathematics, and Computing (CEISMC). He received his Ph.D. in Research, Measurement, and Statistics from the Department of Educational Policy Studies at Georgia State University. Dr. Lingle's research and evaluation work focuses primarily upon STEM education and 21st Century Skill development, such as critical thinking, self-regulated learning, and teamwork. Current evaluation and research activities include K-12 educational curricula and related assessment development, undergraduate engineering students, and school-wide reforms.

Stefanie Wind is an Assistant Professor of Educational Measurement and Evaluation in the Department of Educational Studies in Psychology, Research Methodology, and Counseling at the University of Alabama, where she teaches courses in educational measurement and statistics. She received her PhD in Educational Measurement from Emory University. As a

methodological scholar, her overall research trajectory is aimed at contributing to the development, interpretation, and use of valid, reliable, and fair measures across educational settings. Specifically, her program of research is focused on developing a coherent set of tools based on nonparametric and parametric item response theory (IRT) models for examining psychometric issues within the context of assessments that include rater-assigned scores (i.e., rater-mediated assessments).

Roxanne Moore is currently a Research Engineer at Georgia Institute of Technology with appointments in the G. W. Woodruff School of Mechanical Engineering and the Center for Education Integrating Mathematics, Science, and Computing (CEISMC). Her research focuses on engineering education innovations from K-12 up to the collegiate level. She received her MS and Ph.D. in Mechanical Engineering from Georgia Tech in 2009 and 2012, respectively. She received her BS in Mechanical Engineering from University of Illinois Urbana-Champaign in 2007.