Developing an engineering design process assessment using mixed methods: An illustration with Rasch measurement theory and cognitive interviews

Stefanie A. Wind
The University of Alabama

Meltem Alemdar
Jeremy A. Lingle
Jessica D. Gale
Roxanne A. Moore
Georgia Institute of Technology

Manuscript accepted for publication in *Journal of Applied Measurement.* An earlier version of this manuscript was presented at the annual meeting of the American Educational Research Association in Chicago, IL, April 2015.

Correspondence should be directed to:

Stefanie A. Wind
Assistant Professor of Educational Measurement
Educational Studies in Psychology, Research Methodology, and Counseling
The University of Alabama
313C Carmichael Hall
Phone 205-348-9772
swind@ua.edu

Abstract

Recent reforms in science education worldwide include an emphasis on engineering design as a key component of student proficiency in the Science, Technology, Engineering, and Mathematics disciplines. However, relatively little attention has been directed to the development of psychometrically sound assessments for engineering. This study demonstrates the use of mixed methods to guide the development and revision of K-12 Engineering Design Process (EDP) assessment items. Using results from a middle-school EDP assessment, this study illustrates the combination of quantitative and qualitative techniques to inform item development and revisions. Overall conclusions suggest that the combination of quantitative and qualitative evidence provides an in-depth picture of item quality that can be used to inform the revision and development of EDP assessment items. Researchers and practitioners can use the methods illustrated here to gather validity evidence to support the interpretation and use of new and existing assessments.

Keywords: *Assessment development; Engineering assessment; Mixed methods; Rasch measurement theory*

Developing an engineering design process assessment using mixed methods: An illustration with Rasch measurement theory and cognitive interviews

Recent reforms in science education worldwide include an emphasis on engineering design as a key component of student proficiency in the integrated Science, Technology, Engineering, and Mathematics (STEM) disciplines (e.g., Borgford-Parnell, Deibel, and Atman, 2010; Cardella, Atman, Turns, and Adams, 2008; Kelly, 2014; Kolmos and deGraff, 2014). For example, the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) in the US include engineering design as core idea, and call for "raising engineering design to the same level as scientific inquiry in science classroom instruction at all levels" (p. 1). Current STEM curricula include an emphasis on student proficiency in engineering as a key component of college and career readiness (Auyang, 2004; Carr, Bennett, and Strobel, 2012; Duderstadt, 2008).

Despite the emphasis on engineering design in the development of instructional activities and frameworks for engineering education, relatively little attention has been directed to the development of psychometrically sound assessment methods related to engineering design (Diaz and Cox, 2012). Recognizing this limitation, the Committee on Developing Assessments of Science Proficiency in K-12 has called for the use of systematic, evidence-based approaches to assessment design for the NGSS (National Research Council [NRC], 2014; Pellegrino, DiBello, and Brophy, 2014). This call reflects the current view of validation as an integrated process based on multiple sources of evidence to support the interpretation and use of test scores (AERA, APA, and NCME, 2014).

Purpose

The purpose of this study is to illustrate the use of a mixed-methods technique for gathering validity evidence to guide assessment item revisions within the context of K-12 engineering education. In order to provide context for the methodological demonstration, an illustrative analysis is presented using data collected within the context of a middle-school technology and engineering course. The mixed-methods technique is illustrated using three guiding questions:

1. What does quantitative evidence based on Rasch measurement theory reveal about the psychometric quality of an engineering design assessment?
2. What does qualitative evidence based on cognitive interviews reveal about students' cognitive processing and perceptions of difficulty drivers for items on an engineering design assessment?
3. How can quantitative and qualitative evidence be combined to guide revisions to an engineering design assessment?

The major motivation for combining quantitative and qualitative evidence was to demonstrate a method for converging the two forms of data to bring greater insight to the quality of the assessment items than would be obtained by either type of data separately (Creswell and Plano-Clark, 2011).

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 NGSS (NRC, 2014). Following the recommendations, the theoretical framework for this study

draws upon principles from Construct Modeling[1] (CM; Wilson, 2005) and Evidence-Centered Design (ECD; Almond, Steinberg, and Mislevy, 2002).

Wilson's (2005) CM framework emphasizes the use of evidence obtained from an instrument to make statements about what a student knows and can do based on his/her responses to items. The aspect of CM emphasized in this study is the use of a measurement model that provides validity evidence related to the internal structure and content of an instrument. Essentially, validity evidence related to internal structure describes the degree to which relationships among individual tasks, and the relationship between individual tasks and the total score from an assessment, suggest that the tasks measure a single construct (AERA, et al., 2014).

Similarly, ECD focuses on the role of evidence in developing assessment tasks that elicit a particular construct, the intended inferences from assessment scores, and the nature of the evidence supporting the inferences (Almond, Steinberg, and Mislevy, 2002; NRC, 2014). This study focuses on the evidence model component of ECD, where empirical evidence is used to support the interpretation of responses to assessment tasks as indicators of student achievement with respect to a construct (Mislevy and Haertel, 2006; Snow, et al., 2010).

This study integrates key principles from both CM and ECD to provide a framework for the revision and development of assessment items within the context of engineering education. Specifically, sources of validity evidence for assessment items are 1) quantitative indices of the degree to which assessment items function as expected by a

[1] The term "Construct Modeling" is consistent with the language in the NRC (2014) recommendations for the development of science assessments; this concept is essentially equivalent to the "Constructing Measures" framework presented by Wilson (2005).

measurement model (CM) and 2) qualitative indices of the degree to which individual items reflect the intended construct (ECD).

Methods

Using a sequential mixed-methods design, this study illustrates an approach for developing multiple-choice assessments within the context of K-12 engineering education.

Engineering Design Process (EDP) Assessment

The illustrative analysis in this study focused on an engineering design process (EDP) assessment for middle school students whose items were aligned to engineering concepts in an experimental engineering curriculum project. The instrument used for the illustrative analyses in this study focused on conceptual understanding of the engineering design process rather than a specific design. Specifically, the EDP assessment used in this study included 18 multiple-choice (MC) items developed to reflect one or more stages as well as the overall EDP model used in the curriculum project (see Figure 1).

< Figure 1 here >

Mixed-Methods Design

Figure 2 is a graphical representation of the sequential mixed-methods design (Ivankova, Creswell, and Stick, 2006) used in this study. Phase I was an explanatory design in which results from quantitative data analyses were used to inform case selection for Phase II. Phase II was a convergent parallel design in which quantitative and qualitative strands were used equally and synthesized in order to provide a more-complete understanding of student responses to the EDP assessment items. The quantitative and qualitative data were analyzed in a parallel fashion, with equal priority.

Mixing of the two sources of data occurred after data were collected (at the results stage). Following the notation system proposed by Morse and Niehaus (2009), the design can be represented as follows: (quan →) QUAL + QUAN = More complete understanding.

< Figure 2 here >

Phase I: Explanatory Design (quan →)

Phase I was an explanatory phase that represents a participant-selection variant of the explanatory design used to obtain "initial quantitative results to identify and purposefully select the best participants" (Creswell and Plano-Clark, 2011, p. 86). The first phase of the study included quantitative data collection via the pilot administration of the EDP assessment. The participants were middle school students (6^{th}, 7^{th}, and 8^{th} grade; $N = 415$) enrolled in an experimental engineering curriculum project within two public schools in the US. The course instructors were in their second year of curriculum implementation and taught all three grade levels.

Quantitative data analysis. The quantitative data analysis for Phase I included the use of the dichotomous Rasch model (Rasch, 1960; implemented using Winsteps, Linacre, 2014) to gather quantitative evidence about the psychometric properties of the pre-test administration of the EDP assessment.

***Case selection for qualitative data collection*.** Results from the quantitative data analysis during Phase I were used to create a stratified sample of students to participate in cognitive interviews during Phase II. Based on the distribution of student achievement estimates on the pre-test, students were classified into three groups of approximately equal sample sizes (high, medium, and low achievement levels). Similarly, item difficulty estimates were used to classify items into three subsets (easy, moderate, difficult) based

on the pre-assessment. Six item sets were prepared that included an easy, moderate, and difficult item. As illustrated in Table 1, these results were used to select a sample of 48 students for interviews such that each item was used as an interview stimulus for students with a range of achievement levels. Students were randomly selected for interviews within achievement levels using a blocked design. Of those invited, four students chose not to participate in the interviews; the students who did not participate in interviews are not represented in Table 1. The question sets were presented in the same sequence to students in each achievement level group.

< Table 1 here >

Phase II: Parallel Convergent Design (QUAL + QUAN = More complete understanding)

The second phase of the study included the collection, analysis, and interpretation of qualitative and quantitative data to gain a more complete understanding of student responses to the EDP assessment items.

Quantitative data collection and analysis. During this stage, the post-test was administered to the full sample of students, and the dichotomous Rasch model was used to estimate measures of student achievement and item difficulty. Several diagnostic indices were also examined to evaluate the psychometric quality of the items, including item difficulty estimates, the match between item difficulty and student achievement (targeting), reliability, and data-model fit. These indices describe the degree to which observed item difficulty ordering matches the predicted ordering, the precision of estimates of student achievement, and the degree to which individual items function as expected based on the Rasch model (i.e., fit to the measurement model).

Qualitative data collection. Qualitative evidence was gathered using semi-structured cognitive interviews based on a protocol adapted from previous qualitative and mixed-methods item development studies (DeBoer, et al., 2008; Kaliski, France, Huff, and Thurber, 2011). The interview protocol included a concurrent think-aloud procedure and retrospective probes (e.g., Leighton, 2004), with the EDP assessment items used as stimuli. The semi-structured interview format facilitated in-depth analysis of distractor functioning, sources of knowledge, and the overall clarity of items.

A team of six educational researchers collected the qualitative data for this study. In order to ensure a common procedure, the researchers practiced administering the protocol prior to conducting interviews. The interview protocol was also pilot-tested prior to the implementation of the study.

Qualitative data were collected from 44 students. During the interviews, students were asked to think aloud while responding to assessment items. Subsequently, they were asked to elaborate on their understanding of each item, strategies used to select a correct response, and rationale for eliminating answer choices. Each interview lasted approximately 20 minutes and included approximately three EDP assessment items; this procedure resulted in at least six interviews per item.

***Qualitative data analysis*.** Following Kaliski et al. (2011), an initial qualitative coding framework was specified with four coding categories: A) cognitive processing, B) difficulty drivers, C) test-taking behaviors, and D) miscellaneous. After preliminary analyses, the codes were modified to reflect the scope of responses. Appendix A includes the final coding descriptors within each of these categories, along with examples of

student responses for each code. Data analyses for this study focused on Category A and Category B.

Coding process. Qualitative coding proceeded in three major rounds. During the first round, the authors used the initial framework to code three verbal reports in order to ensure a common procedure. A verbal report for each item is considered a unit of text (e.g., verbal report from Student 1 for Item 1). Each student's interview transcript included a verbal report for about three assessment items/units of text (depending on available time). The same code was not assigned more than once to a unit, but multiple codes could be assigned to a unit. Following the first round of coding, the researchers collaborated to establish agreement and refine the initial framework.

The second round of coding was used to explore the frequency of codes across items. The same process of using codes only once for each unit of text was used so that it was possible to count the number of observations of each code associated with each item without duplication within student. The frequencies were used to identify items that warranted further exploration using both qualitative and quantitative analyses.

A third and final round of coding involved a more-detailed analysis of verbal reports, in which the same code could be applied multiple times. The purpose of this round of coding was to gain a more in-depth understanding of student cognitive strategies for responding to items, implicit and explicit use of the EDP as a reasoning strategy, and perceived difficulty drivers for the assessment items.

Merge and Interpret the Results

Finally, results from the QUAL and QUAN strands of Phase II were merged in order to form a more complete understanding of student responses to the EDP assessment

items and to inform item revisions. For the items that were "flagged" during the quantitative and qualitative analyses, a summary of the results was prepared for further consideration by the research team. Using the results, the research team offered insight into potential explanations for items identified as problematic, along with potential improvements, and revised the items based on quantitative and qualitative evidence.

Results

In this section, results from each stage of the data analysis procedures are described. Because the technique for combining quantitative and qualitative evidence is the primary focus of the study, more detail is provided about Phase II than Phase I. The methods used to synthesize the quantitative and qualitative results from Phase II are illustrated using two items from the EDP assessment.

Phase I (quan →)

Student responses to the pre-administration of the EDP assessment were first examined using the dichotomous Rasch model. This technique was used to support the use of person achievement measures and item difficulty calibrations to inform case selection for qualitative data collection during Phase II. Overall findings from this analysis indicated a range of student achievement measures and item difficulty calibrations on the logit scale. Further, values of Rasch Infit and Outfit data-model fit statistics for students and items were between +2 and -2 for the standardized versions, and within an acceptable range based on critical values for Infit and Outfit *MSE* corrected for sample size based on Smith, Schumacker, and Busch (1998): $1 \pm 2/\sqrt{N}$ for Infit *MSE*; and $1 \pm 6/\sqrt{N}$ for Outfit *MSE.* These results provided evidence to support the use of Rasch model results to inform second stage analyses. Because item revisions were

guided primarily by post-test results, results from the pre-test are not explored in detail here.

Phase II (QUAN + QUAL = more-complete understanding)

Following the post-administration of the EDP assessment, the Rasch model was used to calculate measures of student achievement and item difficulty, along with indicators of data-model fit for items in detail. Results from the post-assessment (rather than the pre-assessment) were used to inform item revisions, as student responses should reflect knowledge and skills from the engineering curricula.

Variable map. Figure 3 is a variable map based on the Rasch model that illustrates the locations of students and items on a common linear scale that represents the construct (the logit scale). The first column is the logit scale. High logit-scale values correspond to higher achievement and more-difficult items, and low logit-scale values correspond to lower achievement and less-difficult items. The second column shows the logit-scale locations for each student. Higher-achieving students are located toward the top of the variable map, and lower-achieving students are located toward the bottom of the variable map. An asterisk (*) is used to represent six students, and a period (.) represents one student. The third column shows the logit-scale locations for each of the MC items. More-difficult items are located toward the top of the variable map, and less-difficult items are located toward the bottom of the variable map.

< Figure 3 here >

When examining the variable map for an assessment, it is useful to consider the match between the location of student and item distributions on the scale that represents the construct. Information about this targeting is used to determine the degree to which

assessment items provide meaningful information about student achievement. More precise information is provided when item difficulty estimates are aligned with student achievement estimates. Figure 3 indicates generally good targeting between student achievement and item difficulty on the post-administration of the EDP assessment. However, there are some high-achieving students whose locations are not matched by any assessment items (between about +2.00 and +3.00 logits). For future administrations of the EDP assessment, new items might be created that are targeted students with these levels of achievement.

Rasch model calibrations. Table 2 summarizes the Rasch model results from the post-administration of the EDP assessment; the results in this table correspond to the variable map shown in Figure 3. The student measures and item calibrations correspond to the locations that are plotted in the variable map (Figure 3). The average student location on the post-test was 0.19 logits, which is higher than the average item location ($M = 0.00$, $SD = 0.15$). Table 2 also includes standard error (*SE*) estimates for students and items. In the context of Rasch measurement theory, *SE* describes the precision for each item location estimate, with smaller values indicating greater precision.

< Table 2 here >

Further, Table 2 includes data-model fit statistics, which provide an index of the degree to which assessment items approximate the expectations of the Rasch model. The results in Table 2 indicate that on average, the data-model fit statistics were within their expected range when data fit the Rasch model, with average values of standardized Infit and Outfit *MSE* around 0.00. Furthermore, average values of the unstandardized fit statistics fall within the recommended range based on Smith, et al.'s (1998)

recommended sample-size-corrected critical value for Infit *MSE* of 0.90 < Infit *MSE* < 1.10 ($1 \pm 2/\sqrt{N} = 1 \pm 2/\sqrt{415} = 1 \pm 0.10$), and Outfit *MSE* of 0.71 < Outfit *MSE* < 1.29 ($1 \pm 6/\sqrt{N} = 1 \pm 6/\sqrt{415} = 1 \pm 0.2$). However, inspection of the individual values of fit statistics for the items on the post-test administration revealed several items whose fit statistics exceed the critical values that suggest acceptable data-model fit.

Finally, Table 2 includes indices of reliability for students and items. Within the context of Rasch measurement theory, reliability is examined using the *Reliability of Separation* statistic (*Rel*) and a chi square statistic (χ^2). When data fit the model, the *Rel* for students can be interpreted in an analogous fashion to coefficient alpha. For items, *Rel* describes the spread of differences in the difficulty to correctly answer each EDP assessment item. In addition, a χ^2 can be calculated for students to determine whether the differences among logit-scale locations are statistically significant. The results from the EDP assessment indicated a moderate *Rel* statistic for students (*Rel* = 0.76) and a high *Rel* statistic for items (*Rel* = 0.96), with significant differences among individual students and items ($p < 0.01$). Taken together, these findings of overall adequate data-model fit for students and items support the interpretation of the variable map as an illustration of student and item locations on the construct measured by the EDP assessment instrument explored in this study.

Results: Qualitative Data Analysis

The qualitative analysis included coding student transcripts. Results from the analysis indicated that all of the items elicited a cognitive strategy related to at least one stage of the EDP, and all items except Item 1 were associated with at least one difficulty driver. The qualitative analysis also included an examination of student responses in

terms of the overall effectiveness of the EDP items in eliciting the stages of the EDP. As illustrated in Table 3, qualitative results indicated a variety of responses that explicitly or implicitly indicated use of each EDP stage.

< Table 3 here >

Merge the Results

Next, the results from each strand were considered together to identify items that warranted revision. Among the 18 items, six were flagged during both the quantitative and qualitative analyses, two were flagged in only the qualitative analyses, and the remaining ten items were not flagged in either analysis. For each item flagged for further analysis by both or either method of analysis, a summary of quantitative and qualitative results was prepared in order to facilitate discussion and guide item revisions. Appendix B includes a sample item summary document for one of the items described below.

Interpret the Results

In this section, two items are presented as examples to illustrate the discussion and revision process. The first item (Item 6) is a standalone item (not scenario based), and the second item (Item 16) is based on a scenario for which several items were included in the EDP assessment.

Item 6

Item 6 was a standalone item that was flagged for review during both the quantitative and qualitative data analyses. The original item was as follows:

6. Which of the following is something that an engineer would NOT do when defining a problem?

   A. Identify what the final design solution needs to be able to do.

B. Think about the cost and materials available for the design.

C. **Propose a design solution that will solve the problem. (Correct response)**

D. Conduct research on technology ideas related to the problem.

Rasch indices of psychometric quality indicated that the item was relatively easy ($\delta = 1.57$ logits, $SE = 0.11$), compared to other items on the assessment. The values of data-model fit statistics exceeded their critical values, suggesting unexpected response patterns associated with this item. A detailed residual analysis for Item 6 indicated that all of the unexpected observations resulted from high-achieving students incorrectly responding to the item when they were expected to provide a correct response.

The item review and revision process also included an examination of the results from student responses to Item 6 during the cognitive interviews. Although the item was intended to elicit the EDP stages of *problem definition* and *problem understanding*, qualitative results for this item indicated that student responses did not elicit these stages. Interestingly, although the item was the easiest during the post-assessment, none of the students correctly responded to the item during the cognitive interviews, and student responses were fairly balanced across the three distractors during the interviews. Two students re-read the item stem and changed their answer when they focused on the use of the word "NOT" in the stem. One student misunderstood the question because they did not focus on the word "NOT," and quickly eliminated the correct answer choice. However, the student's explanation for ruling out this answer choice indicated correct understanding of the problem definition stage of the engineering design process. Four of the eight students who were interviewed selected answer choice D. Students'

explanations for selecting option D indicated that they thought conducting research would be "going off topic" and not focusing on the problem. All of the students referenced the EDP and/or their engineering class activities when describing their strategies for eliminating answer choices.

During the review of the results for Item 6, the research team agreed that a potential cause for the problems associated with this item was the word "NOT" in the stem. Further qualitative analyses revealed that, although option B represented the best answer to the question, the problem definition phase as defined in the curriculum is fairly precise, implying that other choices including D were plausible answers. This was a technical error that was uncovered through the student responses. The group also discussed potential misunderstandings related to answer choice D, which could be construed as a correct or partially correct answer. The group decided to drop the word "NOT," and revise answer choice D. The item was revised as follows:

> Which of the following is something that an engineer would do when defining a problem?
>
> **A. Identify what the final design solution needs to be able to do. (Correct response)**
>
> B. Think about the cost and materials available for the design.
>
> C. Propose a design solution that will solve the problem.
>
> D. Research technology related to the problem.

Item 16

Item 16 was also flagged for review during both the quantitative and qualitative data analyses. This item was based on a scenario that describes a situation in which an

engineer has been hired to design a system for cleaning shopping carts for a grocery store. As originally written, the scenario that served as the stimulus for Item 16 was as follows:

> A grocery store has a problem. Customers have complained that their carts are unsanitary and may be spreading germs. The grocery store contacted an engineering company to design an automatic system for cleaning shopping carts. The system must:
>
> - Cost less than $500
> - Use less than 10 gallons of water per day
> - Clean 100 carts in 30 minutes or less

The original version of Item 16 was as follows:

> 16. On her first try, an engineer finds that the design almost meets all of the criteria the grocery store provided. It is under $500, uses 9 gallons of water per day, and cleans 100 carts in 35 minutes. The engineer decides that she has addressed the problem and will share her design with the grocery store.
>
> What mistake did the engineer make?
>
> A. She did not define the problem and identify criteria and constraints.
>
> **B. She did not test and modify her design to meet the criteria. (Correct response)**
>
> C. She did not try to determine if her design met all the criteria.
>
> D. She used proper procedures and didn't make any mistakes.

The Rasch indices of psychometric quality for Item 16 indicated that the item was relatively easy ($\delta = 1.64$ logits, $SE = 0.12$), compared to other items on the assessment,

and data-model fit statistics suggested unexpected response patterns. A detailed residual analysis indicated that all of the unexpected observations resulted from low-achieving students selecting the correct answer choice when they were expected to answer it incorrectly.

During the cognitive interviews, most of the students described the steps of the EDP generally in the expected order, made some allusion to engineering class or the EDP, and used vocabulary related to the stage of the EDP intended to be elicited by this item: *conceptual design.* However, results from the cognitive interviews indicated several misunderstandings, particularly related to option C. For example, the students who answered incorrectly generally selected C, which suggests confusion between the *conceptual design* and *evaluation* stages of the EDP. One student also stated that he did not understand option C at all, and that was the reason he did not choose it. In addition to confusion related to option C, several students indicated that they were not sure whether the engineer had tested the design before drawing conclusions related to the criteria. Specifically, they were unsure if "On her first try" meant that the engineer conducted a single test, or that the item was describing the first design. Further, almost all of the students who were interviewed about this item indicated a lack of understanding of the term "criteria," noting that they did not remember learning about this word in their engineering class.

During the item review, the research team agreed that the term "criteria" might have resulted in misunderstanding due to the use of "requirements" in the curriculum to describe a similar concept. The curriculum writers suggested replacing "criteria" with "requirements" in order to better match curricular materials. The item stem was also

revised to clarify that the engineer was testing the first design for the cart cleaning system. Additional revisions included shortening sentences in the item stem, and more clearly identify the client in the scenario. The revised scenario and item were as follows:

A grocery store manager has a problem. Customers have complained that their carts are unsanitary and may be spreading germs. The grocery store contacted an engineering company to design an automatic system for cleaning shopping carts. The system must:

- Cost less than $500
- Use less than 10 gallons of water per day
- Clean 100 carts in 30 minutes or less

16. When she tested her first design, an engineer found that the design met <u>almost</u> all of the requirements the grocery store provided. It is under $500, uses 9 gallons of water per day, and cleans 100 carts in 35 minutes. The engineer decided that she solved the problem. Now she will share her design with the grocery store manager.

What mistake did the engineer make?

A. She used proper procedures and did not make any mistakes.
B. She did not try to determine if her design met all the design requirements and constraints.
C. **Even though she tested her design, she did not try to modify her design to meet the requirements. (Correct response)**
D. She did not clearly define the problem with a problem statement and identify requirements and constraints for the design.

Results from the subsequent administration of the revised EDP assessment indicated that the revised versions of Item 6 and Item 16 function as expected by the Rasch model, with acceptable values of data-model fit statistics (Item 6: Infit *MSE* = 1.07, Std. Infit = 1.32, Outfit *MSE* = 1.04, Std. Outfit = 0.55; Item 16: Infit *MSE* = 0.96, Std. Infit = −0.75, Outfit *MSE* = 0.95, Std. Outfit = −0.57).

Summary and Conclusions

As pointed out by Pellegrino (2012), "assessments do not offer a direct pipeline into a student's mind" (p. 833). In order to ensure valid interpretation and use of assessment results, evidence must be collected to determine how students understand questions and select answer choices. In particular, it is essential to collect evidence to support the inference that an assessment instrument is not unduly influenced by threats to validity, including construct underrepresentation and construct-irrelevant variance (AERA, et al., 2014; Douglas and Purzer, 2015). Through the combination of quantitative psychometric results and qualitative evidence related to student cognitive processes, this study illustrated a systematic technique for establishing validity evidence for multiple-choice assessment items within the context of a K-12 engineering assessment.

What does quantitative evidence based on Rasch measurement theory reveal about the psychometric quality of an engineering design assessment?

When the dichotomous Rasch model is applied to student responses to multiple-choice items, and diagnostic statistics and displays can be used as evidence of item quality in terms of difficulty, targeting, reliability, and data-model fit. In the case of the illustrative analysis, quantitative results indicated that the EDP assessment items demonstrated good psychometric properties. However, several items were flagged for

further review based on data-model fit statistics and patterns observed in residual analyses.

What does qualitative evidence based on cognitive interviews reveal about students' cognitive processing and perceptions of difficulty drivers for items on an engineering design assessment?

The use of a qualitative coding framework for cognitive interviews was demonstrated that could be used to identify construct-relevant and construct-irrelevant influences on student responses that warrant further attention during item revisions.

In the case of the illustrative analysis, qualitative results indicated that students used the EDP model as a reasoning strategy when responding to the MC items, and that all of the items elicited skills related to stages of the EDP. Detailed analyses indicated that some items elicited engineering skills or stages other than those intended during item writing, and that some items elicited construct-irrelevant difficulty drivers, such as misunderstanding related to vocabulary and guessing.

As noted above, the third round of analysis included a more in-depth qualitative analysis that focused primarily on identifying the EDP stages described by the students for each of the items after instruction (post-test) in order to ensure that the assessment instrument could be interpreted as a measure of conceptual understanding of the design process. This process was also used to identify under-represented stages for which new items should be included in future iterations of the assessment, as well as to inform revisions to the curriculum. Results from this analysis revealed that, in general, student responses indicated use of one or more stages of the EDP for each of the 18 items.

Discrepancies between the observed cognitive strategies and those intended for each item were used to inform additional revisions to items and revisions to the curriculum.

In terms of item difficulty, the qualitative analyses focused on determining the degree to which items were difficult for students based on their conceptual understanding of the EDP and its application to the assessment items, which could potentially be addressed through revisions to the curriculum, and the degree to which item difficulty was related to construct-irrelevant factors that could be addressed through item revisions. A full discussion of findings from the qualitative analysis in terms of the EDP as a cognitive model is provided in Authors (in press). Future research should also explore the alignment between item difficulty and various stages of the EDP.

How can quantitative and qualitative evidence be combined to guide revisions to an engineering design assessment?

The illustrative analysis in this study demonstrated the synthesis of quantitative and qualitative results to explore the degree to which items are functioning as expected by a psychometric model with useful measurement properties, and the degree to which there is evidence that student response patterns reflect the application of intended cognitive processes, rather than construct-irrelevant factors. Whereas the use of only quantitative indicators, such as item difficulty and data-model fit provides information about the overall functioning of an item, this information does not provide much insight into specific directions for item revisions related to students' cognitive processes, difficulty drivers, test taking strategies, and other observations only available through qualitative analyses. On the other hand, the use of only qualitative indicators of item quality is limited in terms of the scope of information that can be provided due to

practical constraints in data collection and analysis. In the context of the current study, the item review committee was able to use both sources of evidence to identify potential revisions to the EDP items before it was re-administered.

Results from the illustrative analysis indicated that patterns identified using the Rasch-based (quantitative) techniques were confirmed and supplemented by themes identified in the qualitative analyses of student interviews. Further, results from the subsequent administration of the EDP assessment indicated improved item-level results in terms of item-to-student targeting, reliability, and data-model fit. It is beyond the scope of the current study to describe in detail the results from the subsequent administrations of the EDP assessment, which is part of a larger ongoing item bank development project that has included the pilot-testing and revision of many new multiple-choice items following the administration described in this study.

Implications

The illustrative analysis described in this study was used to demonstrate a systematic mixed-methods approach for collecting and examining validity evidence for multiple-choice assessment items in the context of K-12 engineering education. Specifically, the illustration was based on a MC EDP assessment that was aligned with the instructional objectives of a middle school engineering curriculum. The approach illustrated in this study provides a set of techniques that can be used by researchers during the item development and revision stages of the assessment development process, and by practitioners to inform the use of existing instruments in new settings in order to strengthen the validity argument for a multiple-choice assessment (Kane, 1992). The qualitative coding framework presented in this study offers an example of a systematic

method for summarizing student think-aloud responses that provides insight into student conceptualization and application of an EDP. In particular, examination of the alignment between the intended and observed cognitive strategies at the individual item level can not only inform item revisions, but can also highlight areas for additional curricular emphasis or revision aimed at addressing misconceptions or misunderstanding. In domains where limited research and few validated K-12 assessments exist, methods such as those illustrated here are necessary in order to inform the interpretation and use of new and existing assessment instruments.

With the increased attention to the EDP in the K-12 engineering curricula, this study informs the ongoing dialogue about the development of EDP assessments, with an emphasis on the role of validity evidence to support inferences for the interpretation and use of assessment instruments.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Almond, R. G., Steinberg, L. S., and Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5), 1-64.

Auyang, S. Y. (2004). *Engineering—An endless frontier*. Cambridge, MA: Harvard University Press.

Borgford-Parnell, J., Deibel, K., and Atman, C. J. (2010). From engineering design research to engineering pedagogy: Bringing research results directly to the students. *International Journal of Engineering Education*, *26*, 748–759.

Cardella, M., Atman, C. J., Turns, J., and Adams, R. (2008). Students with differing design processes as freshmen: Case studies on change. *International Journal of Engineering Education*, *24*, 246–259.

Carr, R. L., Bennett, L. D., and Strobel, J. (2012). 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), 1-26.

Creswell, J. W., and Plano Clark, V. L. (2011). *Designing and conducting mixed methods researc*h. Los Angeles, CA: Sage.

DeBoer, G. E., Lee, H. S. and Husic, F. (2008). 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* (pp.153-182). New York, NY: Columbia University Teachers College Press.

Diaz, N. V. M., and Cox, M. F. (2012). An overview of the literature: Research in P-12 engineering education. *Advances in Engineering Education*, *3*(2), 1-37.

Douglas, K. A., and Purzer, S. (2015). Validity: Meaning and relevancy in assessment for engineering education research. *Journal of Engineering Education, 104*, 108-118.

Duderstadt, J. 2008. *Engineering for a changing world: A roadmap to the future of engineering practice, research, and education*. Ann Arbor, MI: The Millennium Project, The University of Michigan.

Ivankova, N. V., Creswell, J. W., and Stick, S. S. (2006). Using mixed-methods sequential explanatory design: from theory to practice. *Field Methods*, *18*, 3–20.

Kaliski, P. K., France, M., and Huff, K. (2011, March). *Using think aloud interviews in evidence-centered assessment design for the AP World History exam*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kelly, A. E. (2014). Design-based research in engineering education: Current state and next steps. In A. Johri and B. M. Olds (Eds.) *Handbook of Engineering Education Research* (pp. 497-518). New York, NY: Cambridge University Press.

Kolmos, A., and De Graff, E. (2014). Problem-based and project-based learning in engineering education. In A. Johri and B. M. Olds (Eds.) *Handbook of Engineering Education Research* (pp. 497-518). New York, NY: Cambridge University Press.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6-15.

Linacre, J. M. (2014). Winsteps (Version 3.81.0) [Computer software]. Beaverton, OR: Winsteps.com.

Mislevy, R. J., and Haertel, G. (2006). Implications for evidence centered design for educational assessment, *Educational Measurement: Issues and Practice, 25*, 6-20.

Morse, J. M. and Niehaus, L. (2009). *Mixed methods design: Principles and procedures.* Walnut Creek, CA: Left Coast Press.

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

National Research Council (NRC). (2014). *Developing assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching*. 49, 831–841.

Pellegrino, J. W., DiBello, L. V., and Brophy, S. P. (2014). The science and design of assessment in engineering assessment. In A. Johri and B. M. Olds (Eds.) *Handbook of Engineering Education Research* (pp. 571-600). New York, NY: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, Chicago, IL: University of Chicago Press, 1980).

Smith, R. M., Schumacker, R. E., and Bush, J. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66–78.

Snow, E., Fulkerson, D., Feng, M., Nichols, P., Mislevy, R., and Haertel, G. (2010). *Leveraging evidence-centered design in large-scale test development (Large-Scale Assessment Technical Report 4).* Menlo Park, CA: SRI International.

Wilson, M. R. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Table 1. *Matrix Sampling Technique for Cognitive Interviews*

Item Sets			Student (Pre-test Achievement Level*)																
Set	Item	Pre-test Difficulty*	1 (L)	2 (M)	3 (H)	4 (L)	5 (M)	6 (H)	7 (L)	8 (M)	9 (H)	10 (L)	11 (M)	12 (H)	13 (L)	14 (M)	15 (H)	16 (L)	17 (M)	18 (H)
	1	Easy																		
1	2	Moderate	X	X	X															
	3	Difficult																		
	4	Easy																		
2	5	Moderate				X	X	X												
	6	Difficult																		
	7	Easy																		
3	8	Moderate							X	X	X									
	9	Difficult																		
	10	Easy																		
4	11	Moderate										X	X	X						
	12	Difficult																		
	13	Easy																		
5	14	Moderate													X	X	X			
	15	Difficult																		
	16	Easy																		
6	17	Moderate																X	X	X
	18	Difficult																		

***** The student sample was divided into three groups of approximately equal size based on achievement on the pre-test (Low, Medium, and High). Likewise, the items were divided into three groups based on their difficulty level on the pre-test (Easy, Moderate, Difficult). Each interview focused on one item set, which contained one easy, moderate, and difficult item.

Table 2. *Quantitative Results: Summary Statistics from the dichotomous Rasch model (Po Assessment)*

Logit-Scale Measure	Student	Item
M	0.19	0.00
SD	1.51	0.15
N	415.00	18.00
Infit *MSE*		
M	0.99	0.98
SD	0.24	0.14
Std. Infit *MSE*		
M	0.00	−0.27
SD	0.90	2.42
Outfit MSE		
M	1.01	0.99
SD	0.54	0.29
Std. Outfit *MSE*		
M	0.10	−0.14
SD	0.90	2.65
Separation Statistics		
Reliability of Separation	0.76	0.96
Chi-Square	406.50*	23.40*
Degrees of Freedom	414.00	17.00

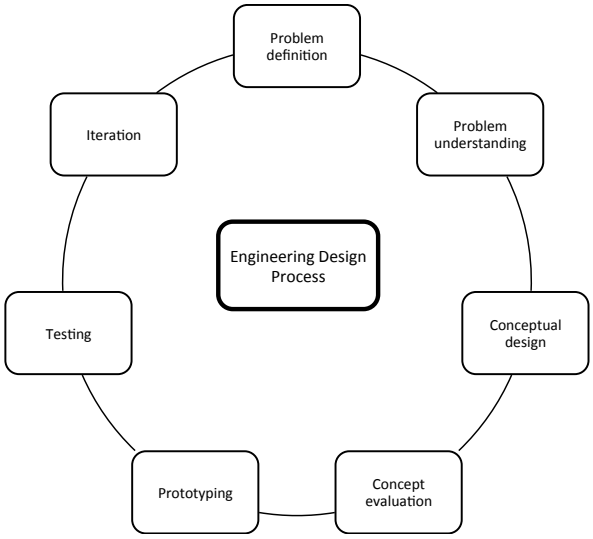
* $p < 0.05$

Table 3. *Summary of observations about EDP as a cognitive model*

EDP Stage	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 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 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 don't hear 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.*

Table 3, continued.

EDP Stage	Types of Responses	Example Student Response
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, and 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.*
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.*



Operational Definitions for Engineering Design Process Stages

Problem Definition	The engineer /designer identifies a specific problem to be solved. The goal during this stage is to clearly identify the need that is to be addressed.
Problem Understanding	The engineer /designer identifies critical aspects of the problem that will affect their success. The engineer/designer should identify the following: 1) Client/market/customer and their requirements and preferences; 2) Functions that the designed solution must carry out; 3) Constraints or resource limitations that will affect the solution; and 4) Engineering requirements for the problem that directly affect customer needs.
Conceptual Design	The engineer/designer identifies possible solutions to the problem at a conceptual level, without focusing on 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	The engineer/designer develops a testable version of a designed solution.
Testing	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.
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.

Figure 1. *Conceptual Model (Subset of Coding Category A)*

Figure 2. *Mixed-methods design*

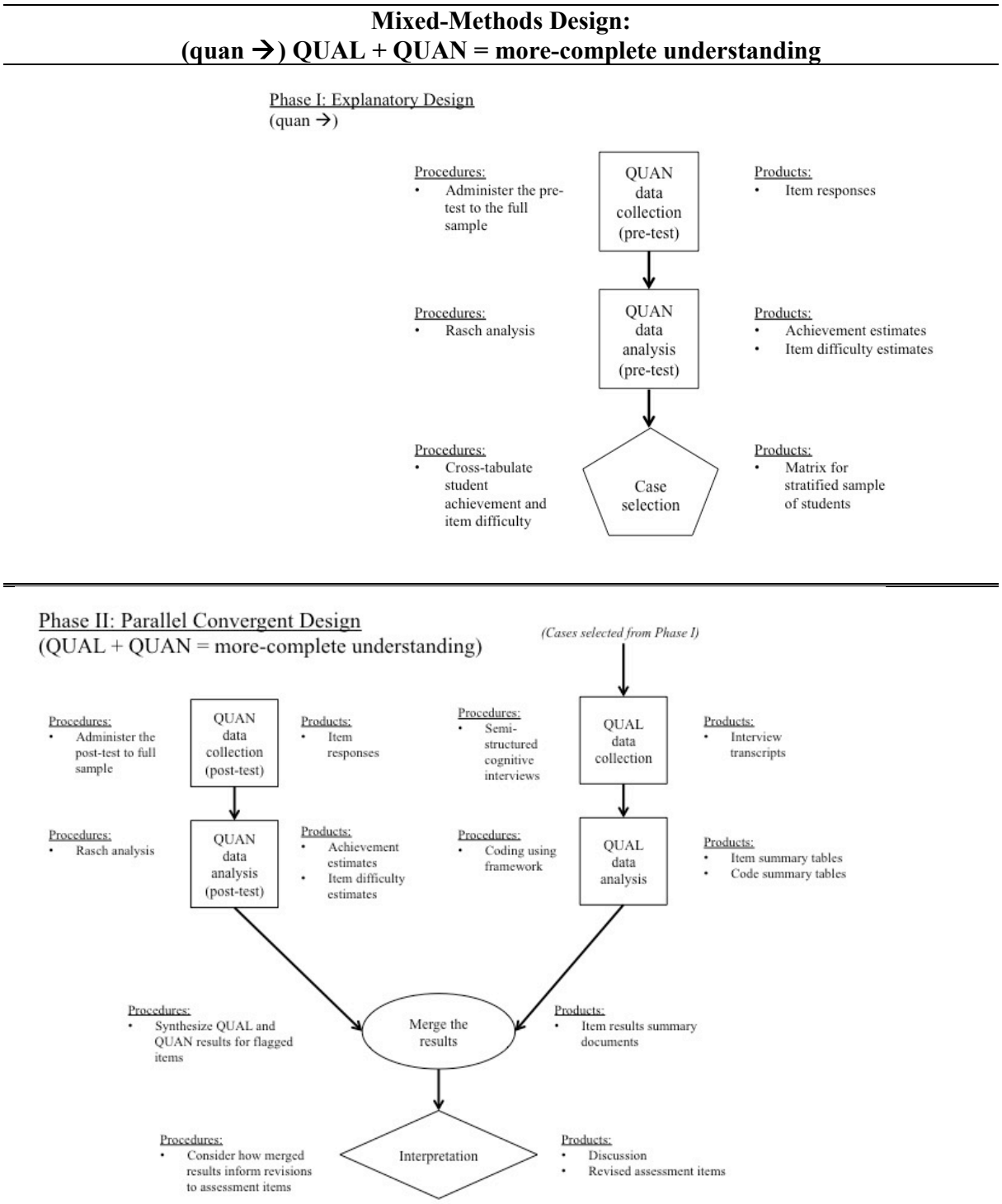
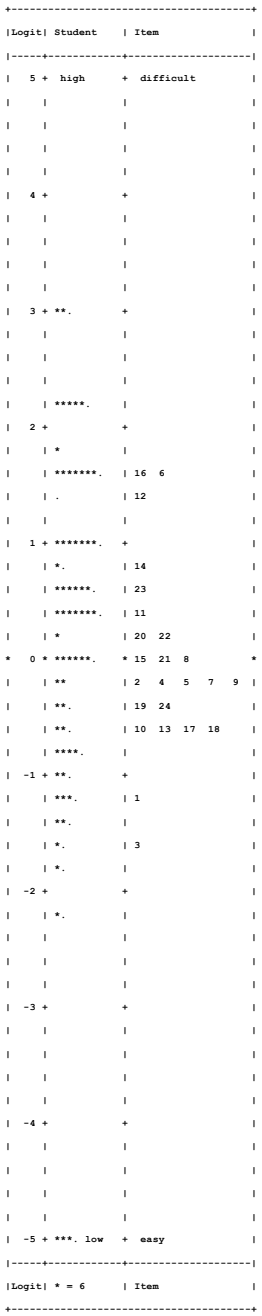


Figure 3. *Quantitative Results: Variable Map (Post-Assessment)*



Appendix A:
Qualitative Coding Framework

Coding Category A: Cognitive Processing

Code	Definition	Example Response
Engineering design process*	• Student demonstrates (explicitly or implicitly) use of one or more specific stage(s) from the engineering design process to answer the question	(See Table 3)
Evidence of intended skills	• Student generally refers to the engineering design process or demonstrates engineering reasoning when answering the question without a specific reference to or demonstration of a stage in the engineering design process.	*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 process.*
Factual recall	• Student uses recall to answer the question rather than applying a specific engineering concept or skill	*Because when [my teacher] said the definition of constraints, I guess it just helped me out with it.*
Guessing	• Student selects the answer at random • Student does not have the knowledge/skills necessary to answer the question	*I sort of just guessed on that one, because that question is confusing.*
Background characteristic	• Student draws upon personal background or experiences to answer the question	*I have had a similar situation, not exactly but my grandpa had to build a container to hold feed for cows but they had to design it right so that it could be kept covered in the winter but also they could fit under and eat. He had to work through it and had to design many different designs to be able to fix it.*

* See Figure 2

Category B: Difficulty Drivers

Code	Definition	Example
Length	• Student indicates that the length of item stimulus material makes an item difficult • Student indicates that the length of the item stem makes it difficult • Student indicates that the length of an answer choice makes it difficult	*I almost picked that because it was too many words and it got confusing.*
Stimulus material (graphics, charts, etc.)	• Student indicates that the stimulus for an item (e.g., graphics, charts, etc.) makes the item difficult • Student indicates that the length of the stimulus makes the item difficult	*I didn't get out what this little thing is, or what it's supposed to be.*
Degree of familiarity	• Students have not had an opportunity to learn the content, making the item less familiar • Students indicate that the item context is unfamiliar	*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.*
Quality of distractors	• Student indicates that some distractors were easy to eliminate • Student indicates that two or more distractors appear to be plausible options	*A and D were sort of kind of alike, so they sort of confused me.*
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 • Student does not understand an answer choice	*I was confused with C because I really didn't understand it.*

Note. These codes are adapted from Kaliski, France, Huff, and Thurber (2011)

Category C: Test-taking behaviors

Code	Definition	Example
Process of Elimination	• Student eliminates one or more answer choices to help arrive at an answer choice	*Well, D doesn't seem right, because it doesn't even have anything to do with racing. B is a truck instead of cars, so it doesn't have anything to do with cars. Let's see, A, repair the engine in a car that will not start ... It doesn't say anything about a car that would not be starting, so I think C would be the right answer because A, B, and D doesn't have anything to do with the question."*
Re-state or Re-read	• Student rephrases or rereads item stimulus material to understand the item context • Student rephrases or rereads the item stem to understand what is being asked • Student rephrases or rereads an answer choice to understand a response option	*I'm trying to see what the question is asking again.*
Misread Question	• Student misreads stimulus material, an item stem, or an answer choice	*Student: After identifying which designs meet the cri ...* *Interviewer: You're close. Criteria, yup.* *Student: ... criteria,*
Change Answer	• Student decides to change their answer	*No, wait. I think it would be A*
Scaffolding within the item	• Student indicates that a specific item detail clued them to the correct answer	*At first, I thought it would be B, but then I saw the word where it said the word ring toss. It doesn't say anything about ring toss in there, so - That was almost likely the game that would choose with their own game in there.*

Note. These codes are adapted from Kaliski, France, Huff, and Thurber (2011)

Coding Category D: Miscellaneous

Code	Definition	Example
Correct Response	• Student selected the correct answer	(Varies by question)
Incorrect Response	• Student selected an incorrect answer	(Varies by question)
Difficulty Thinking Aloud	• Student struggled with the think-aloud task	*I don't talk well when I read out loud. I like to read in my mind. I read faster.*
Researcher Prompt	• The researcher prompted the student for an answer	*Remember to think aloud while you're deciding between the choices.*
Stimulus Material Irrelevance	• The student noted that part of the item stem or a stimulus was irrelevant to the question	*The question is telling you about another requirement that is so the diagram wasn't really helping.*

Note. These codes are adapted from Kaliski, France, Huff, and Thurber (2011)

Appendix B:
Example Item Results Summary Document

Item 6: Quantitative and Qualitative Results Summary

- **Original Item:**

Which of the following is something that an engineer would NOT do when defining a problem?

A. Identify what the final design solution needs to be able to do.
B. Think about the cost and materials available for the design.
C. **Propose a design solution that will solve the problem**.*
D. Conduct research on technology ideas related to the problem.

- **Quantitative results summary:**

Quantitative Index	Observed Value	Accepted/Expected Range
Proportion correct (*p*-value)	0.27	Usually aim for an overall range of 0.30 – 0.80 for an assessment
Difficulty measure (logits)	1.57	Look at overall targeting (match between student achievement and item difficulty distributions)
Standard Error	0.11	Smaller is better
Infit Mean Square Error *(MSE)*	1.22	Adjusted for sample size: 0.90 ≤ Infit *MSE* ≤ 1.10
Standardized Infit	3.97	-2.00 ≤ Std. Infit ≤ 2.00
Outfit *MSE*	1.77	Adjusted for sample size: 0.71 ≤ Outfit *MSE* ≤ 1.29
Standardized Outfit	6.06	-2.00 ≤ Std. Outfit ≤ 2.00

  - **Residual analysis:**

Results from the residual analysis indicated negative residuals for this item. This finding indicates that all of the unexpected observations resulted from high-achieving students incorrectly responding to this item, when they were expected to provide a correct response.

- **Qualitative results summary**

None of the interviewed students selected the correct answer to this item, and their responses were fairly balanced across the three distractors. Two students re-read the item stem and changed their answer when they focused on the use of the word "not" in the stem. One student misunderstood the question because they did not focus on the word "not," and quickly eliminated the correct answer choice. However, the student's explanation for ruling out this answer choice indicated correct understanding of the problem definition stage of the engineering design process. Four of the eight students who were interviewed selected answer choice D. Students' explanations for selecting option D indicated that they thought conducting research would be

"going off topic" and not focusing on the problem. All of the students referenced the engineering design process and or their engineering class activities when describing their strategies for eliminating answer choices.

Summary of Item Revision Discussions from Research Team

The research team agreed that the most pressing problem with this item was the word "NOT" in the stem. Because of the qualitative responses, it was also noticed that although option B represented the best answer to the question, the problem definition phase as defined in the curriculum is fairly precise, implying that other choices including D were plausible answers. This was a technical error that was uncovered through the student responses. We also discussed some issues with answer choice D, which was worded in a way that sounded overly complex. The group decided to drop the word NOT, and revise answer choice D so its intended meaning was clearer.

Revised Item:

Which of the following is something that an engineer would do when defining a problem?

A. Identify what the final design solution needs to be able to do.*
B. Think about the cost and materials available for the design.
C. Propose a design solution that will solve the problem.
D. Research technology related to the problem.