
AC 2012-2987: ENGINEERING CONCEPT ASSESSMENT: DESIGN AND DEVELOPMENT

Dr. Jenny Daugherty, Purdue University, West Lafayette

Jenny Daugherty is an Assistant Professor in the Technology Leadership & Innovation Department at Purdue University.

Dr. Rodney L. Custer, Black Hills State University

Debra Brockway, Stevens Institute of Technology

Daniel A. Spake, Purdue University

Daniel A. Spake is a master's student in the Organizational, Leadership & Supervision Department in the College of Technology at Purdue University. He received his B.S. in technology education from North Carolina State University.

Engineering Concept Assessment: Design & Development

Introduction

While much of the focus on K-12 STEM education has concentrated on science and mathematics, engineering is increasingly seen as a viable educational approach. Some of the reasons include that engineering (a) provides authentic educational problem solving contexts for mathematics and science; (b) may increase the number of students interested in STEM areas, particularly from underrepresented populations¹; and (c) might facilitate the technological literacy of all students². In 2009, the National Academy of Engineering (NAE) convened a Committee on K-12 Engineering Education resulting in a report that stressed the contribution of engineering to the development of an effective and interconnected STEM education system.³ Several engineering-oriented programs have emerged, ranging from those designed to promote technological literacy to those designed to prepare students for post-secondary engineering education.

In order for the impact of engineering at the K-12 level to be understood, there is a need to clearly define the nature and substance of engineering at that level, as well as to perhaps develop a useful and effective set of engineering standards.⁵ For example, in a multiple case study project of five prominent teacher professional development projects, one of the primary findings was a distinct lack of definition of engineering and an inability of teachers to identify their learning outcomes^{6,7}. One of the most alarming aspects of this void was the teachers' inability to reflect on what they were learning related to engineering, apart from a vague understanding of the engineering design process. Without a clear understanding of core engineering concepts, the connection to student learning is tenuous at best. This void also poses serious problems for high quality curriculum and assessment. As the NAE Committee on K-12 Engineering Education observed, a "critical factor is whether teachers—from elementary generalists to middle school and high school specialists—understand basic engineering concepts and are comfortable engaging in, and teaching, engineering design".⁸

This paper describes the design and development of an assessment to measure the learning of engineering concepts. Important issues that have been encountered during the development process are discussed so as to serve to inform future efforts in developing sound instrumentation for K-12 engineering. The assessment development process and resulting issues can be appropriately framed within the context of an engineering design problem. Just as Dym, Little, Orwin, and Spjut (2009) defined engineering design as a "a systematic, intelligent process in which designers generate, evaluate and specify designs for devices, systems or processes whose form(s) and function(s) achieve clients' objectives and users' needs while satisfying a specified set of constraints." (p. 6)⁹, the development of the Engineering Concept Assessment (ECA) has also followed an iterative decision-making process that is important to describe. The criteria for developing the ECA, the constraints that were identified during its development, and the resulting trade-offs are discussed. The primary research question explored in this paper is: how can a valid and reliable assessment be developed to assess the learning of engineering concepts within secondary science?

The Design Problem

The development of the ECA is being informed by a variety of factors including the expertise of the project's team and other experts, the research literature, and a pilot test instrument that was developed to measure a small set of engineering concepts¹⁰. In addition, this design problem is situated within the context of a National Science Foundation Discovery Research K-12 funded project that aims to research the viability of a concept-driven approach to teacher professional development¹. Research is being conducted to understand how science teachers learn engineering concepts and the issues and problems encountered during implementation. However, in order to determine the effectiveness of such an approach, a valid, reliable assessment of engineering concepts is needed. While there are several post-secondary engineering education assessments^{i.e., 11, 12, 13} and some K-12 engineering-oriented instrumentation, most focus on a general awareness of what engineers do or are specific to the content or processes associated with the design context. The design problem for the DR K-12 project is to develop an assessment that can be used to measure the extent to which teachers and high school students understand core engineering knowledge appropriate for secondary level engineering education and their relationship to science content.

An essential input to the design process was the identification of a set of core engineering concepts and some decisions about what concepts (and how many) to include. Two studies have been conducted to identify key concepts^{14, 15} and the National Research Council report, "A Framework for K-12 Science Education" identified cross-cutting engineering concepts important to science¹⁶. Custer, Daugherty, and Meyer's (2010) study provided an in-depth analysis of a broad range of engineering-related literature and conducted focus groups with engineering educators and engineers to identify core engineering concepts. This process resulted in thirteen concepts (analysis, constraints, design, efficiency, experimentation, functionality, innovation, modeling, optimization, prototyping, systems, trade-offs, and visualization). Rossouw, Hacker, and de Vries (2010) conducted a Delphi study and a subsequent panel meeting to generate concepts and contexts that can be used for developing curricula. Three rounds were conducted resulting in five main concepts (design [as a verb], systems, modeling, resources, and values) and 16 sub-concepts (optimising, trade-offs, specifications, invention, product lifecycle, artefacts ['design as a noun'], structure, function, materials, energy, information, sustainability, innovation, risk/failure, social interaction, and technology assessment).

An important initial step to developing the Engineering Concept Assessment was to narrow and refine the list of concepts gleaned from the two studies. Equally important was to develop a systematic process for defining the core concepts. A definition of concepts is essential in order to provide a clear and precise foundation on which to develop assessment items. In order to inform the process for defining the conceptual base from the lists developed by these two studies, a variety of texts were reviewed to identify definitions for the concepts. The texts consulted include introduction to engineering textbooks (used primarily with freshmen engineering students), standards documents, and philosophy of engineering literature. Definitions were documented if they were specific to the engineering domain but in a broad conceptual way (not

¹ This material is based upon work supported by the National Science Foundation under Grant No. 1158615. Any opinions, findings, and conclusions of recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

to a specific engineering discipline). The definitions were recorded verbatim, as well as any supporting text that further elaborated the concept. This information was presented to the project leadership team (5 researchers). Based on this information, the two studies and report cited above, and the need to achieve additional focus for the assessment process, the team decided to focus on a smaller set of primary concepts that are central to engineering, important at the secondary level, and can provide strong links to science education. Four primary concepts emerged and sub-concepts were identified under these concepts serving to highlight key components. The concepts and sub-concepts are:

- Design (constraints, trade-offs, optimization, prototyping)
- Analysis (life-cycle, cost-benefit, risk)
- Systems (structure, functions, interrelationships)
- Modeling (visualization, prototyping, mathematical models)

The design problem was further narrowed to:

- Develop an assessment that can be used to measure the extent to which teachers and students understand design, analysis, systems, and modeling.

The Criteria

The goal of assessment in general is to assign scores to individuals so that some attribute of those individuals is accurately reflected by that score. To obtain an accurate measurement, the criteria for any assessment are that the test scores are reliable and the inferences made on the basis of those scores are valid.¹⁷ Reliability and validity have specialized definitions in the context of assessment. Here, reliability refers to the consistency of the test results over multiple administrations regardless of when the test was administered and scored and who scored it.¹⁸ There are both basic and sophisticated methods for determining reliability.

In the widely used classical theory, reliability is a measure of the internal consistency of an instrument and is calculated by the intercorrelation of test items, or coefficient alpha. The generalizability theory provides a more sophisticated approach to reliability including the consideration of the use of the test scores and is determined by calculating the analysis of variance.¹⁹ Essentially, then, one perspective of reliability is the degree to which the items hang together. And, given that reliability is calculated by the intercorrelation of test items, this aspect of the test is a function of the number of items of which it is composed and the number of test takers. A third and more recent method of determining reliability depends on a more computationally complex model: item response theory (IRT). IRT models provide more precise estimates of test takers' ability, as these models can be displayed in graphical form for each item, showing the probability an individual at any ability level will answer the item correctly. As might be expected, these models require larger sample sizes than methods that rely on traditional measurement theory.²⁰

The second criterion for all assessments is test validity. There are two aspects of validity: the validity of the measurement and the validity of the decisions that are made based on the test results. Validity answers questions such as "Is the test fair?"; "Does the test measure what it purports to measure?"; and "Are the test results useful for the intended purpose?"²¹ Assessing the content validity of a test is not an exact measure; rather it is a judgment regarding the degree to

which it represents a particular content domain. Essentially, the procedure for evaluating content validity involves the following steps:

- Describe the content domain.
- Determine the areas of the content domain that are measured by each test item.
- Compare the structure of the test with the structure of the content domain.²²

A content validity study requires engaging a group of content experts in a review of an elaborated delineation of the content domain and the test items. Following a discussion of the representation of the domain on the assessment, the group makes recommendations for further development of the test.²³

Although this aspect of validity seems straight forward, Messick (1989) and others have observed that it is essentially impossible to prove not only that a test is valid but that a test measures what we purport it measures.²⁴ And validity involves much more than content validity. Four fundamental aspects of validity that are common themes in the writings of psychometricians are:

- Validity is not a property of a test. Rather, it refers to the use of a test for a particular purpose.
- To evaluate the utility and appropriateness of a test for a particular purpose requires multiple sources of evidence.
- If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose.
- Evaluating test validity is not a static, one-time event; it is a continuous process.²⁵

Sireci (2007) is critical of limiting the evidence of test validity to the subjective analysis of test content. He argues that the evidence should also include empirical analysis of test score and item analysis data.

In addition to a reliable and valid instrument, other criteria for the assessment include those that are specific to this situation and are based on the content domain and population being assessed. These criteria are that the assessment items are to address the aforementioned engineering concepts in the context of life and physical science; and that the assessment will be targeted to the level of high school teachers and their students. The content domain can have implications for item types to be included on the assessment. In this case, the content domain (design, analysis, systems, and modeling) suggests a need for some constructed response (CR) item types, in addition to multiple choice (MC) items. This decision should be informed by content and measurement experts.

The criteria for the ECA are:

- A reliable and valid instrument
- That measures understanding of engineering concepts
- Within a life and physical science context
- For high school teachers and students.

The Constraints

As is the case with virtually all engineering design problems, time and cost are two constraints that will significantly impact the scope and design of the assessment and the methods used to

analyze the results. These constraints, along with content and measurement expertise, impact all assessment development, from classroom tests to state-wide assessments, and national or international assessments. Considering two extremes in assessment development with respect to constraints, specifically, teacher-developed assessments for classroom use and the National Assessment of Educational Progress (NAEP), will highlight the range of responses in dealing with these constraints and provide a lens for considering appropriate approaches to addressing these constraints in developing an assessment for an education research project.

High school science teachers administer several tests in any given semester. Teachers new to the profession or to a particular course have limited time to create these tests and generally rely on resources purchased by the school district (e.g. ancillary materials offered by textbook publishers) and other resources that are freely available (e.g. online test banks), in addition to any original items the teacher creates. The reliability and validity of these tests is not determined beyond the judgment of the classroom teacher, who likely has no background knowledge in the area of psychometric methods and therefore may not be aware of issues related to reliability and validity.

At the other extreme is the development of NAEP assessments. A very brief summary of procedures involved in the development, scoring, and analysis of NAEP assessments as they are presented in the technical documentation will illustrate the time and cost investment on the part of the National Assessment Governing Board (NAGB), the National Center for Education Statistics (NCES) and their contractors to produce assessments that are reliable and valid for measuring what students in grades 4, 8, and 12 know and can do in a variety of subjects.²⁶

Development of the framework and specifications for NAEP assessments generally takes approximately 18 months and involves renowned subject area experts on the cutting edge of cognitive research and instructional practice working in committees to draft an initial document. The framework delineates what is to be measured at each grade level in a specific content area and how it is to be measured. The specifications document is a companion document that is the blueprint for developing the assessment. It includes many sample items and is used by the test development contractor to create the assessment. The draft framework is disseminated and public forums are held to obtain input from a wide array of constituents on the content of the framework. This is an iterative process. When the framework and specifications have been approved by NAGB, the assessment contractor will begin developing items for the assessment.

An item pool is generated that consists of more than twice the number of items needed for the assessment. The items in this pool then undergo review by content experts, teachers, and experts on sensitivity and bias. Pilot tests with small groups of students and cognitive studies are also conducted with the items. NCES guides the revision of the items and, in the case of CR items, the corresponding scoring rubrics.

Then a standing committee of content experts, state and local education agency representatives, teachers, parents, and representatives of professional associations reviews the items and rubrics. Each standing committee considers:

- the appropriateness of the items for the particular grade;
- the representative nature of the item set;

- the match of the items with the framework and test specifications; and
- the quality of items and scoring rubrics.

This is followed by a state item review where further feedback is provided. Items are then revised and submitted to the Governing Board assessment development committee for approval prior to field testing. The field test is used to finalize the testing instrument. Items are revised based on student data, and the item set is once again subjected to review following the same procedure described above. Then, a final set of test items is assembled for NCES and Governing Board review and approval.

Following administration of the assessment, CR items are scored and the student scores based on both the MC and CR items are analyzed. Scoring of the CR items is an involved process to assure consistent, valid, and objective scoring. In addition to selecting and training the most qualified individuals for scoring, a subset of the responses for the CR items are double scored to ensure reliability among scorers, and student responses from prior administrations of the assessment are intermixed into the responses for the current year to ensure that consistency is maintained over time. Analysis of test data involves collecting data for individual items to check for reliability and differential functioning among various subgroups of students, among other item-level statistics. Then IRT is used to estimate the measurement characteristics of each item. Finally, reporting scales and a database are created, allowing comparisons of results over time and comparisons between groups of students, for example.

Although the NAEP assessments are considered low-stakes assessments, at least with respect to students, and state assessments are high-stakes assessments, the latter fall between classroom assessments and NAEP with respect to reliability and validity. Time and money are two constraints that play a significant role in shaping the assessment development process. Items on many state assessments are released each year, necessitating on-going development of entire tests, unlike NAEP which is administered less frequently and which releases a subset of items after each administration.

Time and money are certainly constraints for the design and development of the ECA as well. The 3 million dollar project is funded for five years, with the first year as a pilot within which the instrument must be developed and ready to be tested with the pilot test teachers. In addition, the funds constrain the number of teachers that can participate and the amount of data that can be collected. However, much of the emphasis is being placed on meeting the aforementioned criteria (a valid and reliable assessment) given these constraints.

The constraints for the design of the ECA are:

- Time
- Cost
- Content expertise particular to engineering within life and physical science
- Measurement expertise

The Trade-offs

Given the criteria and constraints, the project's leadership team must make trade-offs in the design of the ECA. Trade-offs are the decisions that are made to optimize a particular design solution often by reconciling the criteria and constraints.²⁷ The optimum goal for this project is to obtain an understanding of the extent to which life and physical science teachers understand the set of engineering concepts and associated sub-concepts that were identified for the project. More specifically, since the goal of introducing science teachers to engineering concepts is to facilitate the learning of science content, it is important to embed the assessment of engineering concepts into science-based scenarios and content. A key assumption of the project is that it is important for science teachers to develop a base level conceptual understanding of engineering in order for that understanding to facilitate the learning of science. This is opposed to simply "doing" engineering-types of activities or "hands on" projects without a grounding in conceptual level understandings.

The procedures being followed for developing the ECA will be informed by the Standards for Educational and Psychological Testing and will likely fall between classroom assessments and state assessments with respect to validity and reliability. The process will begin with development of a test framework, a large part of which is the definition of engineering concepts; and specifications delineating the construct to be measured and describing the characteristics of the test items and testing conditions. Item development will be an iterative process beginning with the development of at least 50% more items than required for the final version of the assessment and will include the review of items by the research team and its advisors, cognitive labs with students in the classrooms of the development and pilot year teachers. Classical test theory will be used to analyze item level data and test data.

In terms of the items, in order to meet the criteria of a valid and reliable assessment that can be completed in a reasonable amount of time, a trade-off was made to focus on a smaller set of concepts. As indicated, there were several engineering concepts that have been identified that could serve as the focus of the assessment however our project has decided to focus on design, analysis, systems, and modeling. Another trade-off that will be made is between validity and a combined reliability/cost factor. Considering the construct to be measured, an assessment with high validity would consist of a significant number of constructed response items. This assessment will have a small number of constructed response items; however, because a larger number of this item type would be prohibitively expensive for the project to score and/or significantly decrease the reliability of the scores obtained.

The trade-offs being made in designing the ECA are:

- Reasonable validity and reliability measures (between classroom and state assessment standards)
- Focusing on a subset of engineering concepts
- Number of multiple choice and constructed response items

Conclusion

The focus of the project is to provide engineering-concept based professional development for science teachers. The working hypothesis of the project is that an engineering design based approach to science will enhance the learning of life and physical science content. To test this

assumption, it is important to be able to assess teachers' base level understanding of engineering concepts prior to and throughout the professional development process. The professional development will include a series of experiences specifically designed to enhance their understanding of the core engineering concepts (e.g., engineering case study analyses, infusing engineering concepts into science curriculum modules, analysis of lessons delivered to students in science classes). In order to accurately explore the interrelationship between engineering-oriented professional development activities, science content, and classroom learning, it is essential that the teachers' level of understanding of engineering concepts be accurately assessed. This is what prompted the development of the Engineering Concept Assessment.

Several unique challenges were identified and are being addressed throughout the instrument development process.

1. In order for engineering concepts to be assessed, it was first necessary to identify and clearly define them based on the literature and using "natural language" appropriate for science teachers at the secondary level.
2. Given the applied and process-based nature of engineering, it was necessary to include a variety of item types into the inventory. At an understanding level, a conceptual grasp of engineering concepts can be understood through the use of multiple choice items. At a more sophisticated level, it was necessary to use scenario-based constructed response items. This was for two reasons. First, the scenarios provided a mechanism for connecting engineering conceptual understandings to physical and life science content. Perhaps more important, constructed response items were necessary to examine the inherent and complex interactions among the concepts. For example, it is necessary to understand that it is not possible to "optimize" multiple "constraints" within a "design" situation, which forces engineers to make "trade-offs" throughout the design process.
3. One of the important design constraints for the instrument development was to contextualize the engineering concepts into science. More specifically, if engineering concepts and experiences are to be used to facilitate the learning of science content, then it makes sense to connect and integrate the two as part of the assessment process. Ultimately, it will be useful to track the extent to which an understanding of engineering concepts and science content are interconnected.
4. One of the challenges of the instrument development process is the extent to which it can be used with science teachers and secondary level science students. Pilot and field testing are needed to examine this feature of the instrument. The assumption is that since engineering concepts are relatively new to both teachers and secondary level students and given the emphasis on the use of "natural language" that an instrument can be designed that will work for both teachers and students.

Bibliography

1. Brophy, S., Klein, S., Portsmore, M., & Rogers, C. (2008). Advancing engineering education in P-12 classrooms. *Journal of Engineering Education*, 97(3), 369-388.

2. Erikson, T. L., & Custer, R. L. (2008). Conceptual foundations: Engineering and technology education. In R. L. Custer, & T. L. Erikson. *Engineering and technology education*. 57th Yearbook, Council on Technology Teacher Education. Woodland Hills, CA: Glencoe, pp. 1-12.
3. Katehi, L., Pearson, G., & Feder, M. (Eds.). National Academy of Engineering and National Research Council. *Engineering in K-12 education: Understanding the status and improving the prospects*. Washington, DC: National Academies Press.
4. Chandler, J. Fontenot, A. D., & Tate, D. (2011). Problems associated with a lack of cohesive policy in K-12 pre-college engineering. *Journal of Pre-College Engineering Education Research*, 1(1), 40-48.
5. National Academy of Engineering, 2010. *Standards for K-12 engineering education?* Washington, DC: National Academies Press.
6. Daugherty, J. L. (2009). Engineering professional development design for secondary school teachers: A multiple case study. *Journal of Technology Education*, 21(1), 5-19.
7. Daugherty, J. L., & Custer, R. L. (2010). Secondary level engineering professional development: Content, pedagogy, and challenges. *International Journal of Design and Technology Education*. Retrieved from <http://www.springerlink.com/content/u032u46kl7437602/>
8. Katehi, Pearson, & Feder, p. 71-72.
9. Dym, C. L., Little, P., Orwin, E. J., & Spjut, R. E. (2009). *Engineering design: A project-based approach* (3rd ed.). Hoboken, NJ: Wiley.
10. Merrill, C. P., Custer, R. L., Daugherty, J., Westrick, M., & Yong, Z. (2008). Delivering core engineering concepts to secondary level students. *Journal of Technology Education*, 20(1), 48-64.
11. McGourty, J. (1999). Four strategies to integrate assessment into the engineering educational environment. *Journal of Engineering Education*, 88(4), 391-395.
12. Maskell, D. (1999). Student-based assessment in a multi-disciplinary problem-based learning environment. *Journal of Engineering Education*, 88(2), 237-241.
13. Qualters, D. M., Sheahan, T. C., Mason, E. J., Navick, D. S., & Dixon, M. (2008). Improving learning in first-year engineering courses through interdisciplinary collaborative assessment. *Journal of Engineering Education*, 97(1), 37-45.
14. Custer, R. L., Daugherty, J. L., Meyer, J. P. (2010). Formulating a concept base for secondary level engineering: A review and synthesis. *Journal of Technology Education*, 22(1), 4-21.
15. Rossouw, A., Hacker, M., & de Vries, M. J. (2010). Concepts and contexts in engineering and technology education: An international and interdisciplinary Delphi study. *International Journal of Technology and Design Education*, Retrieved from <http://www.springerlink.com/content/4u32551r6h44kx42/fulltext.pdf>
16. National Research Council (2011). *Framework for next generation science education standards*. Washington: National Academy Press.
17. Murphy, K.R., & Davidshofer, C.O. (2001). *Psychological Testing: Principles and Applications*. Upper Saddle River, NJ: Pearson Education, p. 73.
18. Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research, and Evaluation*. Retrieved online December 19, 2011 at <http://pareonline.net/getvn.asp?v=7&n=10>.
19. Ibid 14, p.118.
20. Crocker, L. (2006). Introduction to measurement theory. In J.L. Green, G. Camilli, & P.B. Elmore. *Handbook of complementary methods in education research*. Washington, DC: AERA, p. 378.
21. Sireci, S.G. (2005). The most frequently unasked questions about testing. In R.P. Phelp. *Defending Standardized Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 111-121.
22. Guion, R.M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement*, 1(1), 1-10.
23. Ibid 14, p. 150.
24. Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
25. Sireci, S.G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481.
26. National Center for Education Statistics (NCES). (2009). NAEP Technical Documentation. Retrieved online December 26, 2011 at <http://nces.ed.gov/nationsreportcard/tdw/>.
27. International Technology and Engineering Association. (2000). *Standards for technological literacy: Content for the study of technology*. Reston, VA: ITEA.

