THE UNIVERSITY OF SOUTH ALABAMA
COLLEGE OF EDUCATION AND PROFESSIONAL STUDIES

KNOWN-GROUPS VALIDITY AND GENERALIZABILITY OF A MEASURE OF ENGINEERING DESIGN

by

Mary F. Hibberts

A Dissertation

Submitted to the Graduate Faculty of the
University of South Alabama
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in

Instructional Design and Development

December 2017

Approved: RBJohn— Date: 10-17-17

Chair of Dissertation Committee: Dr. R. Burke Johnson
10.17-17

Committee Member: Dr. James P. Van Haneghan
Gayle V Davidson-Shivers 10.17 17

Committee Member: Dr. Gayle V. Davidson-Shivers
11-17-17

Committee Member: Dr. Joshua D. Foster
11-17-17

Chair of Department: Dr. James R. Stefurak
Susan P. Santoli 10-18-17

Director of Graduate Studies: Dr. Susan P. Santoli
J Harold Pardue 11-10-17

Dean of the Graduate School: Dr. J. Harold Pardue

KNOWN-GROUPS VALIDITY AND GENERALIZABILITY OF A MEASURE OF ENGINEERING DESIGN

A Dissertation

Submitted to the Graduate Faculty of the
University of South Alabama
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in

Instructional Design and Development

by
Mary F. Hibberts
B. S., Spring Hill College, 2006
M.S., University of South Alabama, 2009

December 2017

ACKNOWLEDGEMENTS

I sincerely thank Dr. Burke Johnson - my dissertation committee chair, advisor, and friend. You have helped me develop academically and professionally, always believed in me, and made me laugh a lot along the way. I want to thank Dr. James Van Haneghan, a committee member who shares, and has inspired, my love of research, evaluation, and statistics. His knowledge and advice were invaluable to this research study. Thank you to Dr. Gayle Davidson-Shivers and Josh Foster for your support and guidance during the dissertation process. Dr. Alan Martinez, thank you for challenging me, writing with me, and encouraging me to finish. Last, I want to thank Adam Hibberts, my patient husband…I am finally finished.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

Hibberts, Mary F., Ph.D., University of South Alabama, December 2017. Known-Groups Validity and Generalizability of a Measure of Engineering Design. Chair of the Committee, R. Burke Johnson, Ph.D.

Numerous reports have increased national awareness of the need to improve K-12 education in science, technology, engineering, and mathematics (STEM) in order to meet the needs of our increasingly technological workforce (e.g., Honey, Pearson, & Schweingruber, 2014; National Academy of Science, 2007; The President's Council of Advisors on Science and Technology, 2010). Engaging Youth through Engineering (EYE) modules were developed to increase interest and proficiency in STEM fields in middle schools in Mobile, Alabama (Harlan, Pruet, Van Haneghan, & Dean, 2014). The modules covered relevant engineering design challenges integrated into existing science and mathematics curricula and focused on the engineering design process. Initial results were promising and showed that the EYE program was affecting students' engineering design performance and attitudes in some areas, when compared to a control group (Harlan, Van Haneghan, Dean, & Pruet, 2015). However, the assessment instruments for used measuring engineering design performance require further validity and reliability research before researchers can be confident in their interpretations based on assessment data. The purpose of this study was to evaluate the psychometric properties of three engineering design performance assessments developed for the EYE initiative.

Known-groups validity was tested by comparing engineering design scores from a middle school data set, collected by Harlan et al. (2015), with scores from two groups of college students (i.e., college freshmen with little to no engineering experience and senior engineering students). As expected, senior engineering students had better engineering design performance than the other groups (measured by the engineering design assessments developed for the EYE program). However, the assessment instruments used to measure engineering design performance yielded inconsistent results when comparing the groups with less engineering experience. There were also inconsistencies in group differences when comparing scores on four dimensions of engineering design (i.e., depth and breadth of thinking, teams and expertise, critical evaluation of a design, and use of data and research).

A generalizability analysis was used to evaluate the reliability of the three assessment instruments completed by the college students. When considering total performance scores, there was enough generalizability across people, independent of rater and form, to suggest the instruments measured a general underlying engineering design construct. Generalizability coefficients were lower and inconsistent when considering each engineering dimension individually. Overall, the data suggest that total scores from the three engineering design assessments yield reliable results but have weak to moderate validity. Recommendations for future research are discusses including revisions to the assessments and scoring criteria to increase reliability for engineering dimensions, conducting a generalizability study with middle school students, and testing the psychometric properties of the assessment instruments with additional populations.

INTRODUCTION

Success in today’s increasingly technological and competitive world requires a different set of skills and knowledge than were required years ago (Razzouk & Shute, 2012). The United States Congress Joint Economic Committee stated that technological skills are becoming more important to employers as technology becomes a more critical component of a variety of industries (U.S. Congress Joint Economic Committee, 2012). As the world becomes increasingly advanced and high tech, the value of our innovators and workforce depends in part on the success of their science, technology, engineering, and mathematics (STEM) education. Interest in and the quality of STEM education is critical for the United States to remain a global leader. The United States must focus on cultivating a STEM-competent workforce to solve incredible problems in areas such as energy, medicine, the environment, and cyber security (PCAST, 2010). Therefore, we must educate children to compete successfully in a global marketplace in which knowledge is one of the most valuable assets.

Despite this demand, there is a lack of available workers in STEM-related fields in the United States. For many companies, there is a shortage of STEM talent and a lack of skills in science and engineering (Deloitte Consulting LLP, Oracle, & the Manufacturing Institute, 2009). In addition, the United States lags behind other nations in STEM education at the elementary and secondary levels and American students lack both proficiency and interest in STEM fields (PCAST, 2010). Our nation needs to foster a strong educational foundation rooted in science and engineering to prepare students to meet the challenges we face today and in the future.

The STEM Movement

Numerous reports have increased national and international awareness of the critical need to increase participation in STEM fields in K-12 in order to meet the STEM-dependent workforce needs of the 21st century (e.g., Commission on Mathematics and Science Education, 2009; Honey et al., 2014; Organisation for Economic Co-operation and Development, 2008; National Academy of Sciences, 2007; Rennie, Venville, & Wallace, 2012). The President’s Council of Advisors on Science and Technology published a report on the need to increase interest and proficiency in STEM fields (PCAST, 2010). In their report, they acknowledge the challenges many schools face while trying to implement successful STEM programs. Schools often lack math and science teachers who are knowledgeable and passionate about the subjects to inspire students’ STEM interest, and teachers often lack the support they need to be successful (e.g., professional development, engaging curricula, and adequate assessments).

The PCAST report provided recommendations and guidance to prepare all students to be STEM proficient, to motivate all students to learn STEM, and to inspire students to pursue STEM careers. The report supported developing state-led standards, training 100,000 great STEM teachers by 2020, incorporating STEM education in and out of the classroom, creating 1,000 new STEM-focused schools by 2020, and building a strong and strategic leadership for STEM education to promote and monitor STEM improvement progress.

To facilitate the paradigm shift, the Next Generation Science Standards (2013) were developed to be a set of performance expectations (i.e., standards) focused on preparing students for college and careers that meet the 21st century needs. The standards focused on coupling practice with content to encourage transfer. The standards did not dictate curriculum; they allowed flexibility in instruction and assessment of the standards. They were developed to prepare high school graduates for the demands of college and careers that require STEM-related skills such as critical thinking, design, and problem solving.

In reaction to the STEM deficit, a national call to increase students' performance in STEM fields, and better defined learning objectives and standards, K-12 engineering initiatives emerged across the country (e.g., Engineering by Design [ITEEA, 2006]; Engineering is Elementary [Museum of Science, 2005]).

Engaging Youth through Engineering

The Engaging Youth through Engineering (EYE) program, introduced in two middle schools in Mobile, Alabama (Harlan, Pruet, et al., 2014), is a prime example of an initiative to integrate engineering into math and science education. The Mobile Area Education Foundation (MAEF) collaborated with business and community leaders, the Mobile Country Public School System, and the University of South Alabama to address K-12 STEM issues that could improve the STEM workforce deficit in the region. Shortly thereafter, the EYE pilot initiative began to incorporate the engineering instruction into existing middle school curricula.

The EYE program comprises applied engineering education modules that fit into existing science curricula in an attempt to increase interest and to improve performance in engineering. Researchers found promising results. When compared with a matched comparison group, students participating in the EYE program had more confidence in applying STEM skills and valued work associated with STEM careers more (Harlan et al., 2015). Students in the EYE program revised flawed design plans more often and were more likely to utilize data and research in design plans than students not enrolled in the EYE program (Van Haneghan, Harlan, & Dean, 2015). As part of the EYE initiative, Harlan, Dean, et al. (2014) developed three engineering design performance assessments and a scoring rubric (see Appendices A and B). However, the assessments required further evidence of validity and reliability to allow for sound interpretation of assessment data.

Purpose of the Study

The purpose of this study was to investigate the psychometric properties of the three engineering assessment instruments developed to measure engineering design performance in association with the EYE program (Harlan, Dean, et al., 2014). The instruments were administered to a broader population with more expertise in engineering to check for known-groups validity and to evaluate the reliability and generalizability of the instruments. Known-groups validity was evaluated to answer the question: Do the assessments differentiate performance based on membership to groups that are expected to differ (e.g., novices vs. experts)? Specifically, scores from the engineering assessments were compared from (1) middle school students enrolled in an engineering curriculum, (2) middle school students not enrolled in an engineering curriculum, (3) college freshmen with little to no engineering experience, and (4) college seniors enrolled in their capstone-engineering course. Varied levels of engineering experience were expected to influence engineering performance scores across groups supporting the validity of the assessment instruments.

In addition, the reliability and generalizability of the instruments were tested by evaluating the consistency and dependability of scores within participants across assessment instruments. The three assessments were designed to be parallel instruments. The questions on the three assessments were essentially the same; the only difference being tailoring to specific engineering scenarios (e.g., mechanical engineering and civil engineering scenarios). Therefore, individuals who completed all three assessment instruments were expected to score approximately the same on each of the assessments. Evaluating the reliability and validity of these assessments is an important step toward developing and implementing a comprehensive engineering program into math and science curricula.

Research Questions

Four research questions drove the research:

RQ1. Is there enough generalizability across people, independent of rater and form, to suggest an underlying general engineering design construct measured by the assessment instruments?

RQ2. Is there enough generalizability across people, independent of rater and form, on each engineering dimension to suggest four general underlying constructs measured by the assessment instruments?

RQ3. Does engineering experience affect overall engineering design performance?

RQ4. Does engineering experience affect engineering design performance on the following engineering dimensions: depth and breadth of thinking, teams and expertise, critical evaluation of a design, and use of data and research?

Significance of the Study

The EYE program was designed, developed, and implemented as part of a growing initiative to improve students' proficiency and interest in STEM fields to meet the needs of the current workforce. As part of the EYE initiative in middle schools in Mobile, Alabama, Harlan, Pruet, et al. (2014) and Van Haneghan et al. (2015) conducted preliminary research on the validity and reliability of the assessment instruments. The EYE assessments, however, have been tested only with samples of middle school students enrolled in the EYE program and middle school students not enrolled in the EYE program.

Preliminary analyses were promising and indicated that some engineering design skills improved through participation in the EYE curriculum and that the assessments instruments had moderate to substantial agreement interrater reliability. Specifically, Van Haneghan et al. (2015) found that eighth grade students who had participated in the EYE modules for three consecutive years performed significantly better than the comparison group overall and on three individual engineering design dimensions: (a) depth and breadth of thinking; (b) critical evaluation of a design; and (c) use of data and research.

In a longitudinal study, Harlan et al. (2015) tracked two cohorts of students from sixth to eighth grade at participating EYE schools and matched comparison schools over four years. One cohort began with draft versions of the EYE modules that were concurrently implemented, evaluated, and revised. The other cohort began the following year and participated in the finalized EYE modules from grades six to eight. Harlan et al. (2015) also found that students in the EYE program were more confident in their ability to use STEM skills, more knowledgeable about what engineers do, scored higher on some probability and data interpretation items in standardized tests, mentioned the importance of teamwork more, and scored better on some aspects of the engineering design assessment instruments than the control group. There were some inconsistencies in how students performed on the individual engineering dimensions and performance often varied by gender, ethnicity, and cohort.

In a previous study, Harlan, Dean, et al. (2014) found interrater reliability evidence for the associated scoring rubrics (see Appendix B) with Cohen's Kappas ranging from .64 to .83 and rater agreement ranging from 80% to 90%. However, the researchers acknowledged the need for additional investigation to validate the assessment forms using a wider population and to evaluate reliability of the instruments with a more powerful research design. In response, this study extended the existing research to include data from a wider, more advanced population (i.e., college freshmen and advanced college engineering students) to ensure that the assessment instruments are sensitive to the increases in engineering skills and the performance expected with more advanced engineering training. Collecting data on all three assessments from every college participant allowed for further investigation of the reliability of the assessment instruments. The goal of this research was to contribute to the continued design and development of the EYE initiative and associated engineering performance assessments through assessment validity and reliability evidence.

Relevance to Instructional Design

Developing assessments to measure skills and knowledge is a critical step in the systematic design of instruction (Dick, Carey, & Carey, 2009). However, not all assessments add valuable information to an instructional system. Evaluating the validity and reliability of assessments is essential to ensure that the instruments appropriately measure the construct. Testing the validity of an instrument is important because it allows researchers and consumers to trust the conclusions drawn from the results of the assessments and to ensure that the assessments are reliable to measure the intended performance consistently (Kane, 2013).

This study contributes to the body of literature on validation and reliability of open-ended assessments and contributes to the continued improvement of engineering assessments associated with the EYE program.

Limitations

This study is based on non-experimental research. Therefore, the results were interpreted for prediction and only potential causation. A convenience/purposive sample was used rather than a random sample, which limits the statistical generalizations of the results to any known population (Hibberts, Johnson, & Hudson, 2012).

Definitions of Key Terms

Engaging Youth through Engineering (EYE) – A middle school engineering curriculum offered through the Mobile Area Education Foundation (MAEF) in Mobile, Alabama. The EYE curriculum consists of seven classroom-based instructional units that integrate STEM concepts through applied engineering design tasks.

Engineering – systematic and cyclical approach to designing solutions to meet human needs and wants (National Center for Education Statistics, 2014).

Engineering Design – A systematic process involving problem definition, research, design and development, evaluation, and often the redesign of solutions to meet human needs (Dym, Agogino, Eris, Frey, & Leifer, 2005).

Engineering habits of mind – Harlan et al. (2015) identify four engineering habits of mind: depth and breadth of thinking, teams and expertise, critical evaluation of design, and use of data and research. Similarly, the National Academy of Engineering (2008) describes six engineering habits of mind: systems thinking, creativity, optimism, teamwork, communication, and attention to ethical considerations.

Generalizability theory (G theory; Generalizability is also abbreviated as G in G coefficient) – A framework for examining the dependability, or reliability, of measurement instruments. G theory provides reliability indices in the form of G coefficients and isolates individual sources of error such as variance attributed to rater and form (Shavelson, Webb, & Rowley, 1989).

Known-groups validity – A criterion for measuring the validity of a test. If a test has known-groups validity, the test discriminates among groups that are theoretically expected to differ (Cronbach & Meehl, 1955).

Next Generation Science Standards (NGSS, 2013) – A set of standards that outline what students should know and be able to do related to science education. The standards were developed by and for educators and school leaders to help educators design classroom instruction that facilitates and inspires students' interest and skill in STEM.

Reliability – The extent to which a measurement instrument provides consistent results when used over and over again to measure the same thing (Rossi, Lipsey, & Freeman, 2004).

Rubric – A scaled set of criteria that define how performance is categorized into scores. Rubrics include descriptions of each what performance looks like at each performance level to facilitate consistent scoring (Wolf, 1992).

Technology – Any modification to the world to fulfill the needs and desires of humans (National Center for Education Statistics, 2014). Technology often requires the use of science and engineering to invent useful things to solve problems.

Validity – How accurately a measurement instrument measures what it is intended to measure (Vogt & Johnson, 2011).

LITERATURE REVIEW

This chapter begins with a review of the current literature on engineering design curricula in the schools. Next, an engineering design supplemental curriculum and associated assessment instruments are discussed, including the theoretical rationale for their development (Harlan, Pruet, et al., 2014; Van Haneghan et al., 2015). Finally, the reviewed literature is tied to the proposed research regarding the validity and reliability of the engineering design assessment instruments.

Engineering in K-12 Education

Educational standards and assessment have become central vehicles for change in education nationally and at the state level. The National Research Council has been the leader of developing science standards since the 1990s. Their latest report, *A Framework for K-12 Science Education*, described a new generation of science standards with a focus on improving science education (NGSS, 2013). The standards requested a collective national shift to develop instruction that stimulates and builds interest in STEM. In an attempt to increase students' interest and performance in STEM fields, and to meet 21st century workforce needs, an increasing number of reports is raising awareness for the need to transform K-12 education to properly prepare students for the STEM-dependent workforce (e.g., National Academy of Science, 2007; PCAST, 2010). In response, developments geared toward the integration of engineering into existing school curricula have increased (e.g., Honey et al., 2014; Pinnell et al., 2013; Museum of Science, 2005; Van Haneghan et al., 2015).

To facilitate the paradigm shift, the NGSS K-12 science standards (2013) were developed as a set of performance expectations focused on preparing students for science courses in college and STEM professions that meet the needs of the 21st century. The standards focus on coupling practice with content to facilitate transfer to the real world but do not dictate curriculum. They were developed to allow flexibility in how educators design instruction and assessment of STEM and to prepare students for the demands of STEM college courses and careers requiring science-based skills such as critical thinking, design, and inquiry based problem solving.

According to the NGSS (2013), the goal of the Middle School Engineering Design Dimension is for middle school students to define problems, to consider multiple solutions, and to optimize the final design. The framework recommends that students explicitly learn how apply the engineering design process to develop solutions. According to the NGSS (2013), "learning science depends not only on the accumulation of facts and concepts, but also on the development of an identity as a competent learner of science with motivation and interest to learn more" (p. 286). Table 1 shows the four engineering design standards defined in the NGSS. Instructional designers use standards to define the learning outcomes and develop assessments. According to the NGSS, students who demonstrate an understanding of engineering design are able to define the problem, evaluate multiple design solutions, analyze data, and develop a model for testing and modification.

Table 1

Performance Expectations for Engineering Design Understanding

	NGSS Performance Expectation
MS-ETS1-1.	Define the criteria and constraints of a design problem with sufficient precision to ensure a successful solution, taking into account relevant scientific principles and potential impacts on people and the natural environment that may limit possible solutions.
MS-ETS1-2	Evaluate competing design solutions using a systematic process to determine how well they meet the criteria and constraints of the problem.
MS-ETS1-3	Analyze data from tests to determine similarities and differences among several design solutions to identify the best characteristics of each that can be combined into a new solution to better meet the criteria for success.
MS-ETS1-4	Develop a model to generate data for iterative testing and modification of a proposed object, tool, or process such that an optimal design can be achieved.

Many educational initiatives have been introduced across the country in response to the growing need for increased interest and competence in STEM fields and, in particular, engineering design such as Engineering is Elementary (Museum of Science, 2005), Engaging Youth through Engineering (Harlan, Pruet, et al., 2014; Van Haneghan et al., 2015), and Project Lead the Way (2005). A key characteristic of these engineering design curricula and curriculum supplements is the applied nature of the classroom learning activities and associated assessments.

Design-Based Learning for Engineering

Design-based learning is used as an instructional method for engineering skills. It includes authentic, hands-on, and often ill-defined design tasks resembling the activities that normally occur while working within a community of engineers (Puente, Van Eijck, & Jochems, 2013). Increasing the design aspect of STEM education and integrating engineering into existing science curricula aims to develop more lateral thinking skills, to learn to better handle ambiguity, and to develop open-ended problem solving capabilities (Mullins, Atman, & Shuman, 1999).

Design is a central component of engineering and engineering curricula, but it is difficult to teach and it is challenging to develop valid and reliable assessments (Cardella et al., 2011; Dym et al., 2005). Validation of standardized assessment instruments for design skills has significantly lagged behind the development of the instruction (Van Haneghan et al., 2015). In particular, there has been a lack of assessments designed to address the applied nature of the desired design knowledge and skills. Engineering design skills focus on the ability to apply concepts learned during instruction in a variety of situations. Therefore, it is difficult to assess design skills through a set of standardized problems especially across multiple years of a curriculum.

The EYE engineering design assessments were developed as a series of open-ended responses to questions about an engineering design problem (Harlan, Dean, et al., 2014; Harlan, Pruet, et al., 2014; Van Haneghan et al., 2015). Decisions about the assessment instruments and scoring were influenced greatly by the research and theory described in the following sections.

Theoretical Rationale for the EYE Assessments

The EYE modules and assessments were developed in association with the EYE and influenced by the work of Bailey and Szabo (2006). Bailey and Szabo chose to develop an assessment of engineering design process knowledge using open-ended design questions and a rubric scoring method. They chose this assessment method over alternative options (e.g., closed ended questions, final design reports, portfolios of student work, or videos of team design processes) because it met the majority of their assessment criteria described in the following section.

Assessing Engineering

According to Bailey and Szabo (2006), the key criteria for developing an assessment strategy for engineering design process knowledge are that the strategy is (a) at the individual, not team, level, (b) process-focused (i.e., not only focused on the end result), (c) not too time intensive, (d) reliable from student to student, year to year, and problem to problem, and (e) the strategy should span multiple levels of Bloom's taxonomy. The researchers found that the open-ended question strategy met all the criteria. They concluded this strategy was beneficial in assessing higher-order thinking skills such as engineering design performance. However, they anticipated three possible negatives. First, the responses to open-ended questions could potentially pose both inter- and intra-reliability issues. Second, it could be time intensive to score the assessments. Third, the researchers anticipated difficulty developing questions that effectively assess the design skills.

Bailey and Szabo (2006) chose to use an analytic rubric (i.e., rather than a holistic rubric) to provide a more objective way of assessing student responses, to help identify elements that students excel or struggle with, and to enhance instruction while maximizing learning. An analytic rubric consists of a list of ideal points, major traits, and elements that make up an ideal student response with points assigned to each element. A holistic rubric is more subjective and assesses an overall impression of student responses. Ultimately, Bailey and Szabo developed an instrument to measure students' understanding of the design process that required students to critique a completed design proposal depicted in a Gantt chart. Influenced by the work of Bailey and Szabo (2006), Harlan, Pruet, et al. (2014) chose to utilize open-ended questions to measure engineering design performance, to incorporate a critique of an existing engineering design, and to provide an analytic rubric for reliable scoring (see Appendix B).

The EYE was also influenced by Bransford and Schwartz's (1999) perspective on learning and transfer. Most educators would agree that the goal of education is to provide learning activities that have a lasting and positive effect outside of the exact conditions and context in which some skill or knowledge was taught and acquired. Educators hope that what is learned in the classroom at any given time will be applied to similar and novel situations, across courses, over time, and in the real world. Therefore, education for transfer typically requires a mindset of educating people broadly rather than simply teaching individuals to solve specific problems (Bransford & Schwartz, 1999).

Transfer and application of learning is an important consideration in instructional design. Instructional designers strive to create learning environments that allow learners to meaningfully interact with the instructional material to facilitate knowledge construction and integration rather than simple memorization. Considering transfer during the development of instruction and assessments can lead to opportunities for learners to integrate and understand the information in a meaningful way. Meaningful learning allows learners to use their new knowledge in novel situations, rather than just remembering the information (Mayer, 1999).

Measuring Transfer

Measurement of learning outcomes can differ greatly depending on the type of assessment used to evaluate the learning. For example, two learning experiences may result in similar testing outcomes when evaluated in terms of memory but have very different results if they are assessed in terms of transfer rather than memory (Michael, Klee, Bransford, & Warren, 1993). Gagne's (1972) instructional theory includes nine instructional events that lead up to the final event: enhancing retention and transfer. Effective instructional programs focus on the performances used on the job or in the real world in addition to retention (as cited in Reiser & Dempsey, 2007). Therefore, transfer should be considered throughout the instructional design process and influence the performance objectives, instructional strategies, assessments, and evaluation.

Bransford and Schwartz (1999) and Broudy (1977) both argue that transfer is often underrepresented but, in fact, is not rare. Often, inadequate testing affects our ability to recognize transfer when it happens. In assessment, Bransford and Schwartz view transfer as preparation for future learning. Preparation for future learning predicts that even if the exact strategies do not transfer directly to a novel situation, one's experiences will impact how one deals with subsequent experiences.

Bransford and Schwartz (1999) conducted a study that supported the preparation for future learning hypothesis. The researchers asked fifth graders and college students to develop a plan to prevent bald eagles from becoming extinct. They found that the quality of the plans developed by the fifth graders and college students was rather poor. However, students approached the problem differently and asked different types of questions while developing a solution. College students relied on past experiences to guide the types of questions used during problem solving. For example, the college students asked more relevant questions related to a basic understanding of eagle ecology and recognized the importance of gathering more ecological information to help inform their solutions. The younger students asked questions about the eagles themselves (e.g., "how big are the eagles; what do they eat?") rather than about the interdependence of the eagles and their habitat (e.g. "what type of ecosystem supports eagles; what about predators of eagles and eagle babies?").

Bransford and Schwartz (1999) attributed the more advanced questions from the older students to transfer of previously learned biological concerns and general considerations they were exposed to in previous biology courses. While none of the students (college students or fifth graders) came up with good solutions, the college students were able to transfer problem solving strategies from previous experiences to ask more thoughtful and relevant questions. Similarly, Mullins et al. (1999) found that engineering students differed in terms of the sophistication of their design process as a result of continued training and education, but they did not differ in terms of the quality of the end product.

Therefore, the EYE curriculum and assessments were developed to teach and assess engineering "habits of mind." Engineering habits of mind include systemic thinking, creativity, teamwork, communication, and ethical considerations (Katehi, Pearson, & Feder, 2009). The assessments associated with the EYE program evaluate the quality of the design questions and responses similar to the strategy used by Bransford and Schwartz (1999). While the assessments do not relate directly to specific EYE modules, they require design skills and engineering habits of mind to be transferred more generally.

The Influence of Engineering Experience and Engineering Design Skills

Other researchers have found similar evidence of transfer while studying engineering design skills. For example, in the first of a series of studies designed to determine if educational experience influences the sophistication of engineering design processes, Atman and Bursic (1996) used two design questions to compare five freshmen who had just finished reading a design text and five freshmen who had not read the design text. The two design problems were open-ended in nature. The first problem required students to design an apparatus to shoot ping-pong balls at a target some distance away. The second required students to design a solution for a pedestrian and traffic congestion problem on a college campus. The researchers compared the design processes as well as the final solutions. They found that after reading a short excerpt from an engineering design textbook, students spent more time on the design problems and were more sophisticated in their problem solving strategies compared to the freshmen who had not read the design text.

A second study conducted by Mullins et al. (1999) compared 16 freshmen who had completed one college semester to 16 freshmen who had not yet begun college studies. They found that after completing only one semester of a freshman level engineering course, students showed more sophistication in their design processes.

Atman, Chimka, Bursic, and Nachtmann (1999) compared freshmen and senior engineering students as they designed a playground. The researchers found that seniors collected more information, considered more solutions, moved back and forth between steps in the design process more frequently, and produced higher quality designs than the freshmen engineering students. The results from this study support the hypothesis that engineering education influences engineering design performance.

In a follow-up study, Atman, Cardella, Turns, and Adams (2005) collected verbal protocols from 61 senior engineering students and 32 freshman engineering students as they worked out two design problems (i.e., the ping-pong ball and traffic design problems previously described). Verbal protocols are data generated by individuals talking about their current cognitive processes while completing a task (Fonteyn, Kuipers, & Grobe, 1993). Eighteen of the participating students were within-subjects participants contributing first as freshmen and then years later as senior engineering students. The results showed that the solution quality was higher for seniors than freshmen. Seniors also considered solutions, and moved back and forth between design steps more often than the less experienced freshman engineering students.

Atman et al. (2007) conducted a study that compared the engineering design processes of college freshmen, college seniors, and expert engineers through verbal protocols collected while designing a playground. They found that experts spent more time solving the problem, spent more time scoping the problem (i.e., viewing the problem from a variety of perspectives), spent more time on each individual stage of problem solving, and collected more information than the less experienced groups. They found that students with more engineering design experience used these strategies more effectively and that depth and breadth of thinking varied by educational experience. Their results revealed that “problem scoping” and “information gathering” were two design areas that transferred to novel design problems.

In review, Mullins et al. (1999) found that design and problem solving processes improved after one semester of a freshman introductory course in engineering. In a related study, Atman and Bursic (1996) found that simply reading a short piece of text on the engineering design process had measurable effects on students’ design processes. Therefore, in the current study, I expected engineering design performance to vary as a function of engineering design education and experience. Specifically, higher engineering design scores were expected for engineering students than non-engineering students and higher scores were expected for EYE participants than for non-EYE participants.

As another example of design skill transfer, Gruenther, Bailey, Wilson, Plucker, and Hashmi (2009) conducted a study with college seniors before and after completing a capstone course in engineering design. The researchers were interested in the kinds of experiences that increase design knowledge and what that knowledge entails. They measured students’ ability in seven areas of design knowledge: identifying needs, generating ideas, analysis, development and testing, how they proceed through the design process, time allotments, and documentation. Using a rubric, Gruenther et al. (2009) assessed students’ ability in the seven areas of design while students critiqued a completed design proposal. The results indicated that the experience from taking a capstone design course increased ability to identify needs, improved the design layout, and increased relative time allotments of different design activities. Further, the experience of taking the capstone course removed pretest group performance differences present between students that had prior experience with industrial design and those that did not.

Bailey (2008) also used student critiques of completed designs as a measure of engineering design knowledge. In a study comparing undergraduate engineering students and practicing engineers, Bailey investigated differences in their knowledge of the role of problem definition and idea generation. Students' design critiques were evaluated using a scoring rubric. Students learned a significant amount of idea generation skills during a college introduction to engineering course. The researchers found that significant learning related to problem solving occurred only after a senior capstone design course. Experience gained as a practicing engineer also improved performance. These results suggest that (a) critique of a completed engineering design may be used to measure design skills and knowledge and (b) engineering design knowledge and skills change with college education experience and professional experience.

One consistent theme throughout the research on engineering design is that individuals with more advanced engineering knowledge and skills approach the problems differently from those with less engineering experience. Expert designers clarify requirements, actively search for relevant information, summarize and prioritize information and requirements, and consider multiple solutions (Fricke, 1999). Ahmed, Wallace, and Blessing (2003) studied the design differences between expert and novice engineers. They found that the novice engineers tended to focus on trial and error as the primary strategy to generate, implement, modify, and evaluate designs. Experts, on the other hand, tended to make preliminary evaluations of their tentative designs before implementing them and before making full evaluations of tested solutions. In other words, the experienced designers employed integrated design strategies.

According to Honey et al. (2014), many STEM education programs have failed to clearly define and assess STEM learning outcomes and, as a result, there is no way to determine if the instructional goals were met. The NGSS (2013) emphasize the importance of assessments to measure the outcomes of instruction. Instructional designers should consider using multiple forms of assessment that align with the instructional goals and performance standards. The EYE assessments align with the movement from assessing factual and concrete knowledge to assessing interactions between content areas that require an integrated and deep understanding of the material (Honey et al., 2014).

EYE Assessment Instruments

Influenced by the design, transfer, and engineering studies just described, Harlan, Dean, et al. (2014) began developing assessments for the middle school EYE curriculum with the following theory: Given a set of problems with engineering design as the common thread, students would be able to draw on relevant experiences with design to approach the engineering design problems more competently than students without such experiences.

Harlan, Dean, et al. (2014) developed engineering design performance assessments as part of the research associated with the EYE program delivered by the Mobile Area Education Foundation. The EYE engineering curriculum integrates STEM concepts and is a supplemental component of the math and science curriculum offered by Mobile, Alabama middle schools. The seven engineering modules varied in specific engineering content (e.g., mechanical engineering, environmental engineering, and genetic engineering) but consistently focus on the engineering design process. See Table 2 for EYE module titles and the associated engineering fields covered during those modules.

Table 2

EYE Modules

Grade	Design Module	Engineering Field
6	Harnessing the Wind	Mechanical Engineering
6	Don’t Go with the Flow	Environmental Engineering
7	EYE on Mars	Biological Engineering
7	To Puppies & Beyond	Genetic Engineering
7	Catch Me if You Can	Biomedical Engineering
8	Let’s Get Moving	Mechanical Engineering
8	Eco-Friendly Plastics	Materials Engineering

The instructional modules for middle grade math and science instructors were incorporated into existing math and science curricula. Each of the modules included opportunities for students to solve hand-on, real-world design problems related to STEM content. The modules were designed to bridge the gap between state-mandated educational requirements and the needs of potential STEM employers (e.g., innovative problem-solving skills, communication, teamwork skills). The intention was that through repeated engineering education across content areas built into math and science curricula the students could start to develop “engineering design habits of mind” that could be transferred to novel problems.

The assessment of content and skills learned throughout these modules was developed based on the work of Bailey and Szabo (2006) on evaluating design processes and incorporated the depth and breadth of thinking dimension related to design-based problems (Atman et al., 2007). Students answered a series of questions related to a design problem (e.g., a civil engineering scenario related to a trash accumulation problem in a river caused by rainstorms).

Questions asked what the students would need to think about as they considered the problem and whom they would want on their design team. Other questions required students to critically evaluate a proposed solution and describe how they would use and gather relevant research and data (see Appendix A).

The EYE curriculum was developed to align closely with the recommended performance and learning outcomes used to evaluate post-secondary engineering schools and colleges from the Accreditation Board for Engineering and Technology (2000):

- Apply STEM knowledge using the engineering design process.
- Analyze and interpret a variety of data.
- Identify, formulate, and solve problems.
- Communicate effectively.
- Function as part of a multidisciplinary team.
- Use the techniques, skills, and tools necessary in the modern workforce.
- Recognize the need for, and engage in, ongoing learning.

Assessments play a critical role in all instructional systems because they document if students have learned and are able to execute the performance objectives used to design the instruction and assessments (Pellegrino, Wilson, Koenig, & Beatty, 2014). Although, the engineering performance expectations in the NGSS (2013) were published after the development of the EYE modules and assessments, the EYE instruction and assessments support the NGSS (2013) and the NGSS assessment guidelines (Pellegrino et al., 2014).

The EYE modules included both curricular and extra-curricular activities and strategies to promote the learning outcomes in K-12 education. There are four key characteristics associated with each of the EYE modules. Each module (a) addresses an engineering design challenge related to issues from the National Academy of Engineering's (NAE) *Grand Challenges for Engineering* (2008), (b) develops "engineering habits of mind," (c) incorporates technologies and other resources to keep middle school students engaged, and (d) strengthens math and science understanding.

The EYE program was built on the theoretical foundation of the four components of the Bransford, Brown, and Cocking (2000) "How People Learn" framework. Bransford et al. (2000) suggested that instruction should be learner centered, knowledge centered, assessment centered, and occur within communities. Harlan, Pruet, et al. (2014) drew from theoretical concepts discussed in the work of Atman et al. (2007), Bailey and Szabo (2006), Bransford and Schwartz (1999), and Gruenther et al. (2009) when developing the assessments. Specifically, idea generation and the depth and breadth of knowledge were included in the assessments as important components of engineering design (Atman et al., 2007; Bransford & Schwartz, 1999; Gruenther et al., 2009). The EYE assessments also included a critical evaluation of a proposed design as a measure of engineering performance in the EYE assessments (Bailey & Szabo, 2006).

In addition to these components, Harlan, Pruet, et al. (2014) added two components to their assessment criteria. First, they added a dimension in which students identify types of expertise and ideal team composition for effective engineering design. This teaming component is similar to the problem-scoping component identified by Atman et al. (2007) as well as the question generation component. Second, and also in line with Atman and colleagues' problem scoping component, Harlan, Pruet, et al. (2014) added the need to consider the role of research and data in the design process by requiring students to identify the usefulness of data and the types of data that would improve the design process.

These two additional components, taken together with the dimensions identified through prior research, led to four dimensions included in the engineering design assessments:

- depth and breadth of thinking about the problem,
- identification of the team skills and expertise needed,
- critical evaluation of another's application of the design process, and
- ability to use and interpret data to solve an engineering design problem.

Next, the researchers developed a rubric for analyzing scores on the assessments based on the four intended outcomes of the instructional modules and the assessment criteria (see Appendix B).

Scoring the EYE Assessments

Harlan, Dean, et al. (2014) developed grading levels (ranging from zero to three) for responses to questions related to the four engineering dimensions just described. The lowest level, zero, was awarded for irrelevant responses or lack of response to a question. Level one responses addressed the general problem, but failed to connect individual problem components or to address the problem as a whole. Level two responses addressed the problem holistically and demonstrated an understanding of the problem's complexity, but failed to integrate and apply engineering design principles. Level three responses demonstrated an ability to combine and apply engineering design principles. The rubric included descriptive phrases of each level of each dimension to help raters grade consistently. Raters were provided with engineering design process training and verbal examples of the types of responses that might appear at each level.

Psychometric Properties of the EYE Assessments

Harlan, Dean, et al. (2014) tested the rubric for clarity and interrater reliability with two raters scoring the same 30 assessments (15 treatment and 15 control). The researchers evaluated interrater agreement with Cohen's Kappas ranging from .64 to .84 and interrater percent agreement ranging from 80% to 90%, suggesting moderate to substantial agreement. The researchers also captured nuances and trends present in written responses that were not direct outcomes of the rubric.

Van Haneghan et al. (2015) conducted a study to compare eighth grade middle school students' engineering design performance from schools engaged in the EYE with a control sample from similar schools not engaged in the EYE. Students in the study had been at the EYE school for their entire middle school career. The engineering design assessments were scored using the rubric to aid in scoring consistency. The researchers' rubric resulted in Cohen's Kappa for interrater reliability of .85 for depth and breadth of thinking, .85 for expertise needed for the design team, .66 for critical evaluation of the design process, and .79 for use and interpretation of the relevant graphs and data.

After controlling for prior math and reach achievement, Van Haneghan et al. (2015) found that EYE students had higher overall scores than the control group and scored higher than the control group on three of the four design dimensions: (a) depth of thinking, (b) critique of the design, and (c) use of data and research. These results support the hypothesis that the EYE program creates experiences that transfer to novel situations (e.g., the engineering design assessment).

In a follow-up study, Harlan et al. (2015) compared students' attitudes about STEM, STEM related standardized test scores, and performance on the EYE assessment instruments. The study included data from several cohorts of students; all students in the EYE schools had been at the same school from sixth to eighth grade. The researchers controlled for prior math and reading achievement scores. While there were inconsistencies across cohorts in some areas, there was consistent evidence of the EYE students demonstrating engineering "habits of mind" across grades on the engineering assessments and EYE students scored better than the control group on STEM-related standardized test items (Harlan et al., 2015). The researchers also found moderate to substantial interrater reliability of the assessments when raters used the scoring rubric with Cohen's Kappa ranging from .66 to .85 across the four engineering dimensions (Harlan et al., 2015).

Van Haneghan et al. (2014) and Harlan et al. (2015) found evidence supporting the reliability of their assessment instrument and evidence supporting the effectiveness of the instruction. However, they identified the need for a generalizability analysis of the assessment instrument to explore the impact of facets such as form and rater variability. Determining the reliability and validity of assessment instruments is an ongoing process that requires collection of a variety of data. In response, this study continued to investigate the reliability and validity of the engineering design assessments.

Instrument Validation

Alternative assessments (e.g., performance assessments, open-ended questions, essays, portfolios, computer simulations of real-world problems) are considered more authentic forms of assessment and better indicators of higher-order thinking skills and complex reasoning skills than traditional multiple-choice forms of assessment (Archibald & Newman, 1988; Linn, Baker, & Dunbar, 1991; Messick, 1994; Shepard, 1991). Advances in cognitive and social psychology have expanded the range of purposes, contexts of use, forms of activity, and types of assessments used to evaluate higher-order thinking skills. As a result, the number of validity considerations for advanced assessment techniques has increased (Mislevy, 2016). Challenges arise in determining assessment design methods and in choosing validation strategies to measure higher-order skills (e.g., engineering design skills) when using open-ended problem-solving assessments (Mislevy, 2016).

Some argue that performance assessments require a unique set of validity criteria (Linn et al., 1991; Mislevy, 2016) while others argue that the validity concerns of performance assessments are, for the most part, consistent and are even less extensive than the general validity standards (Messick, 1994). Either way, the validity, reliability, fairness, and generalizability of any type of assessment must be evaluated to ensure that appropriate inferences can be extracted from collected results.

Validity is the degree to which the interpretations of scores for proposed uses are supported by evidence and theory (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Messick (1994) described the essence of assessment design as a construct-centered approach. The construct-centered approach to assessment identifies relevant tasks of the construct as well as the rational development of rubrics and scoring criteria based on the construct. Focusing on the construct helps to reduce construct irrelevant variance that can reduce validity. This type of assessment development allows researchers to use assessment scores as evidence of how a population would perform in other situations and is especially useful when assessing complex cognitive processes (Pellegrino et al., 2014).

Therefore, validation requires evaluation of the proposed interpretations and uses of the results of the assessment, and involves collecting multiple forms of evidence to support valid instrument operation. Messick (1989) argues that there are not “types of validity,” rather the different forms of evidence complement and supplement each other. Some important forms of validity evidence include those related to internal/structural validity, such as the relationships among responses to tasks, items, or parts of a test. Other sources of evidence are related to the external structure of an assessment such as the relationship of the assessment scores to similar measures and other background variables. Other potential sources of validity evidence include testing assessments over time or across groups or settings. One can add validity support by investigating variation of test scores as a function of instruction, experience, or as the result of an experimental manipulation (Messick, 1989).

Construct-irrelevant task variance can threaten the validity of an instrument. Validity is threatened when the assessment is too broad and when assessment scores are influenced by variance associated with other distinct constructs such as rater, sampling, items, and form differences (Messick, 1995). Assessing generalized skills, such as engineering design performance, using open-ended responses poses a unique validity problem. Delivering complex tasks to diverse populations can lead to low generalizability in performance tasks. The challenge to attain useful inferences from such an assessment is addressed by evaluating the validity and generalizability of the instrument (Mislevy, 2016).

Generalizability Theory

Generalizability theory (also known as G theory) provides a flexible framework to examine the reliability of assessment instruments. It extends classical theory by estimating the amount of measurement error accounted for by different sources (e.g., rater or form) and provides reliability coefficients tailored to the proposed uses of the measurement instrument (Shavelson et al., 1989). Results from generalizability studies reveal the extent to which results from performance assessments can be verified and generalized to the general construct being measured. Linn et al. (1991) recommended that, at a minimum, researchers collect information on the amount of variability resulting from rater differences and task sampling. This is a necessary condition when considering the generalizability of specific items on an alternative assessment instrument in terms of transfer to the broader domain of achievement that the instruments seek to measure.

Generalizability theory is used to break down the error term, present in most behavioral and psychological measurements, into specific sources of error. Because engineering design performance cannot be measured directly, it is important to identify measurement variance and its sources. For example, applied assessments may vary as a result of individual differences on the construct being measured, or as a result of using different raters, measuring at different times, using multiple instruments, or measuring under different conditions. The different sources of measurement variance must be investigated to isolate measurement of the target construct and to determine the reliability of an instrument. G theory is useful to determine both the consistency and generalizability of results, especially with applied assessment (Briesch, Swaminathan, Welsh, & Chafouleas, 2014).

According to G theory, assessments include the influences of factors outside of the factor of interest. Extraneous factors can influence a participant’s responses to a given assessment on any given occasion. These unrelated factors (e.g., administrator effects) comprise variability, or error. In G theory, these sources of error are called facets. Generalizability studies (e.g., Briesch et al., 2014) that examine multiple problems and explore the impact of facets (e.g., rater variability) are useful in evaluating how reliably an instrument measures differences in performance in different situations (e.g., engineering design skills measured at different times with different assessments).

In this study, generalizability analyses of the engineering assessments were conducted to further break down the total error variance into specific sources of variance.

Specifically, G theory was used to identify the extent to which scores were influenced by form (i.e., the three different scenarios presented in the assessments) and rater. In addition, this study included generalizability analyses of scores within each engineering dimension, and across scenarios. These facets can potentially influence assessment scores, and affect the desired "true" measurement of engineering design task performance.

Bailey and Szabo (2006) identified the key criteria for engineering assessments. One key criterion was that engineering assessments must be reliable from student to student, year to year, and from problem to problem. In order to evaluate the generalizability coefficients and verify that the assessments are reliable across forms and raters, I administered the three assessments to all of the college student participants. The scores from the three assessments were used to evaluate the reliability of the assessment through within-subjects analysis based on G theory. I expected students to generalize engineering design habits of mind to result in consistent scores on the assessment instruments.

Known-Groups Validity

If a measurement instrument is "valid," the instrument produces scores that lead to appropriate inferences about the measured construct. One criterion for measuring the validity of an assessment instrument is known-groups validity. Known-groups validity analyses evaluate whether test scores discriminate among groups that are theoretically expected to differ (Cronbach & Meehl, 1955; Hattie & Cooksey, 1984). For example, if an instrument designed to measure expertise does not distinguish between novices and experts, then there is a problem with the validity of the instrument (Cook, 2014). A review of 417 studies found known-groups validity to be the most commonly used source of construct validity evidence (Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013). Known-groups validity was used in 73% of the studies included in their review. Therefore, in the current study, comparing engineering design assessment scores by level of engineering design experience was used to evaluate the validity of the instruments.

Chapter Summary

In this chapter, a review of the literature on engineering in K-12 education, engineering experience research, and assessment validity and reliability was presented. The hypothesis that engineering experience will affect engineering design performance is supported by the references cited, as are the methods for evaluating validity and reliability in the current study.

METHODOLOGY

This chapter describes the research design, procedures, and data analyses used to evaluate the generalizability, reliability, and validity of the EYE engineering design assessment instruments. The study was designed to build upon research conducted by Harlan, Dean, et al. (2014), Harlan, Pruet, et al. (2014), Harlan et al. (2015), and Van Haneghan et al. (2015) with middle school students by evaluating the validity and reliability of the assessment instruments they developed to assess engineering design performance and engineering habits of mind with a more experienced population.

The EYE assessment instruments were administered to a sample of college students with varied levels of engineering experience and their scores were compared to existing responses from middle-school students (Harlan et al., 2015; Van Haneghan et al., 2015). These data were used to identify evidence of known-groups validity, test the generalizability of the instruments, and demonstrate that the assessment instruments are sensitive to varying levels of engineering design performance.

Construct Validity Theory and Research Questions

The study was designed to determine the validity of the Van Haneghan et al. (2015) engineering design instruments. This process was focused on using known-groups validity and G theory to determine if the instruments operated, as they should, as predicted by the construct validity theory. The research questions and specific hypotheses were as follows:

RQ1: Is there enough generalizability across people, independent of rater and form, to suggest an underlying general engineering design construct measured by the assessment instruments?

RQ1 Hypothesis: The G coefficient for the instruments is high indicating that variability in assessment scores is primarily due to individual differences in engineering design performance rather than item or rater variability.

RQ 2: Is there enough generalizability across people, independent of rater and form, on each engineering dimension to suggest the following dimensions are general underlying constructs measured by the assessment instruments: depth and breadth of thinking, teams and expertise, critical evaluation of a design, and use of data and research?

RQ2 Hypothesis: The G coefficients for the engineering dimensions are high indicating that variability in dimension scores is due primarily to individual differences in engineering dimension performance rather than to item or rater variability.

RQ3: Does engineering experience affect engineering design performance?

RQ3 Hypothesis: Engineering design assessment scores vary as a function of engineering experience. Engineering students score the highest, followed by non-engineering freshmen college students, middle school students enrolled in the EYE program, and middle school students not enrolled in the EYE program, respectively.

RQ 4: Does engineering experience affect engineering performance on the following engineering dimensions: depth and breadth of thinking, teams and expertise, critical evaluation of a design, and use of data and research?

RQ4 Hypothesis: Engineering dimensional scores vary as a function of engineering experience. Engineering students score the highest, followed by non-engineering freshmen college students, middle school students enrolled in the EYE program, and middle school students not enrolled in the EYE program, respectively.

Participants

Twenty-three engineering students and 24 general undergraduate students enrolled at the University of South Alabama participated in the study. Senior engineering students (i.e., students in the final stages of their college engineering curriculum) agreed to participate as an optional activity for their current engineering course. The general education undergraduate group was sampled from students enrolled in an introduction to psychology undergraduate course (i.e., students with little or no college engineering experience). Students in the general undergraduate group volunteered to participate in the study as partial fulfillment of a research participation requirement for the course. All participants read an informed consent form (see Appendix C) before agreeing to participate. The informed consent form described the nature and purpose of the study, ensured participant confidentiality, iterated participants' right to drop out of the study at any point, and discussed compensation for participation in the research study.

Instruments and Operational Measures of Variables

The EYE assessment instruments and corresponding scoring rubrics developed by Harlan, Dean, et al. (2014) were the primary instruments in this study. The design assessments measured engineering design performance. Questions on each of the three assessments were similar; they varied only to suit three engineering design scenarios (see Appendix A).

The design scenarios in the assessments aligned with the engineering focus of the EYE modules for each grade (see Table 2 on page 24). The first scenario was developed for sixth grade students to match the focus of the EYE modules they completed (i.e., mechanical and environmental engineering) by addressing storm-generated trash in a tidal river (referred to as the Dog River assessment). The seventh grade design scenario was to make biofuel out of algae (referred to as the Algae assessment). Eighth grade students addressed a design scenario to modify seatbelts to decrease force-related injuries in elderly adults (referred to as the Seatbelt assessment).

After a brief description of the design scenario, students answered nine questions that measured the four engineering design dimensions. The first two questions prompted students to generate initial questions about the problem. The next two questions asked students to think about the kinds of people they would want on their design team and to describe the team skills and expertise needed to solve the problem. The next three questions followed an example of a design process and solution completed by others. Participants critiqued the design process another team used and offered suggestions to improve the design process. The last two questions required students to make inferences from relevant graphs and charts and to identify additional data and research that would be helpful to design a solution.

Engineering Design Performance

The dependent variable for the study was engineering design performance as measured by scores, ranging from 0 – 12, on the assessments (see Appendix A). The instruments measured four dimensions of engineering design: (1) depth and breadth of thinking, (2) teams and expertise (3) critical evaluation of design, and (4) use of data and research. Together, the scores on the four dimensions make up the dependent variable – engineering design performance, or design habits of mind. Harlan, Pruet, et al. (2014) and Van Haneghan et al. (2015) hypothesized that exposing students to engineering design challenges throughout the middle school years through the EYE program would engender systemic thinking and more insightful questions and responses on the assessments.

Interrater correlations and Cohen’s Kappas for the three assessments showed moderate agreement and were much lower than previously reported by Harlan, Dean, et al. (2014). There were strong positive correlations between raters’ scores on each of the three assessments but low Cohen’s Kappa coefficients indicated better than chance, but not high, agreement between raters (see Table 3).

Table 3

Interrater Reliability Indices for Three Engineering Assessment Instruments

	Dog River	Seat Belt	Algae
Cohen’s κ	.26*	.38*	.16*
Pearson’s *r*	.89*	.88*	.86*

Note. *p < .01

Engineering Design Experience

The independent variable for research questions three and four was engineering experience. Engineering experience was categorized as four levels based on educational experience in engineering: (1) middle school students not enrolled in EYE schools, (2) middle school students participating in the EYE program, (3) college freshmen with little to no college engineering experience (i.e., general education undergraduate course), and

(4) college seniors enrolled in capstone engineering design courses. Harlan et al. (2015) and Van Haneghan et al. (2015) collected the middle school student data from 2009 – 2014 and provided the data set for use in this study.

In addition to the engineering design assessments, the college students completed a brief questionnaire (see Appendix D) to collect demographic information and information about the students' past experience and interest in engineering design (e.g., previous engineering courses completed, design projects, extra-curricular activities related to STEM fields). Responses to the questionnaire were used to examine group characteristics related to engineering experience.

Procedure

The Institutional Review Board at the University of South Alabama approved this study prior to participant recruitment and data collection (see Appendix E). College participants were recruited during their normally scheduled classes. Data collection followed a brief introduction to the study and distribution of informed consent forms.

All college-student participants completed all three of the engineering design assessments and the demographic and engineering design experience questionnaire. Participants were randomly assigned to one of six counterbalanced assessment orders so each assessment appeared in each possible ordering position approximately an equal number of times. Completion of the three assessments took 90 minutes or less and was completed in a university computer lab using Survey Monkey®. Two trained raters scored the assessments using the grading rubric developed by Harlan, Dean, et al. (2014), see Appendix B. The rubric facilitated scoring of engineering design skills in terms of four dimensions.

Data Analysis

Generalizability and Reliability

A generalizability analysis was conducted to examine the facets that influence assessment scores in addition to the measurement of engineering design performance. Data from the two college student groups were used for the analysis to allow for a fully crossed design. The middle school students did not complete all three assessment instruments and, therefore, were not included in the analysis. Scores from both raters on the three assessments were included in the analysis to evaluate the reliability of the assessment through within-subjects analysis based on G theory. I expected students to generalize design habits of mind across the three assessments and exhibit cross-scenario consistent scores for engineering design performance.

The fully crossed, person by form by rater design ($p \times f \times r$) potentially provides unique information to indentify previously unaccounted variance sources attributable to rater and form differences. The $p \times f \times r$ design can be broken down into seven effects (see Table 4).

Table 4

Components of Variance for a p × f× r Design

Person (*p*)	Universe score variance for the object of measurement
Form (*f*)	Constant effect for all persons due to form differences
Rater (*r*)	Constant effect for all persons due to score differences from rater to rater
Person-form (*pf*)	Variation in person ability from one form to the next
Person-rater (*pr*)	Variation in person ability from one rater to the next
Form-rater (*fr*)	Constant effect for all persons; variation in forms by rater
Residual (*pfr, e*)	Residual consisting of the interaction of p, f, and r, and/or random events

Note. Adapted from Zaidi, Swoboda, Kelcey, and Manuel (2017).

The G theory analyses explored the impact of facets to determine whether there is sufficient reliability in measuring differences in engineering design skill across situations. Looking at these facets provides a more comprehensive account of the reliability of the instruments than simply calculating Cronbach’s alpha or a test-retest correlation because G theory exposes more specific sources of error within the instruments (Mushquash & O’Connor, 2006). Five G theory analyses were conducted. One analysis evaluated rater and form variance for the total scores on the three assessment instruments. The remaining four analyses looked at rater and form variance for scores on the four dimensions of engineering design performance, individually.

Person (*p*) is included as a facet in all generalizability analyses. Ideally, differences between individuals’ scores are the main source of variance. If the largest variance component is attributable to person, the results indicate that engineering experience explains the majority of engineering design performance variability. Generalizability analysis output also provides the interactions between the variables, person, rater and form.

Form (f) was included as a facet to see if the three different assessment instruments influence performance or if they can be considered parallel forms used to measure engineering design performance. Form as a facet is another way to determine alternate-form reliability.

Rater (r) was included as a facet to determine whether meaningful differences existed between scores from different individuals. This analysis reveals how closely a single rater's score represents the average rating of all possible raters who could have possibly scored the assessment. The generalizability output shows how much error variance is attributed to differences in the raters scores compared to variance attributed to individual differences in performance and form error.

G theory was used to examine the relative effects of raters, form, and person on score reliability. G theory was used to estimate variance components involved in assessment. In this study, the univariate design ($p \times f \times r$) for the total engineering design scores engineering dimension scores were examined.

Engineering Experience and Known Groups Validity

Three one-way ANOVAs (i.e., one for each engineering assessment) were used to compare assessment scores for the four different levels of engineering expertise: (1) middle school students enrolled in the EYE, (2) middle school students in control schools, (3) general education undergraduate students, and (4) senior engineering students enrolled in their capstone design course. Engineering design performance was the dependent variable, as measure by scores on the three engineering design assessments. I expected the results of the ANOVAs to support the hypothesis that level of engineering experience influences engineering design performance by yielding statistically significant main effects for the independent variable of engineering experience.

Specifically, I expected to find statistically significant differences between the mean scores where middle school students not enrolled in EYE would score the lowest, middle school EYE students would score higher than those not enrolled in EYE, college freshman would score better than middle school students, and senior engineering students would score higher than all other study participants would. These results would provide evidence of known-groups validity for the assessment instruments and would support the hypothesis that the assessment instruments are sensitive to skill variability based on engineering design experience. These results would also add support to the findings presented in the literature review – engineering experience affects engineering design performance (e.g., Atman & Bursic, 1996; Mullins et al., 1999).

Three additional ANOVAs were used to determine the effect of engineering experience on engineering design performance for each of the four engineering dimensions. I used three 4 × 4 mixed ANOVAs, for which the first independent variable was a between-subjects variable with four levels of engineering experience and the second independent variable was a within-subjects variable with the four levels of engineering dimensions. These ANOVAs were used to determine if there were differences in peoples' performance on the engineering design dimensions based on how much engineering experience they had.

Chapter Summary

This chapter outlines the methods used to evaluate the effect of engineering experience on engineering design performance and the psychometric properties of the EYE assessment instruments. The participants were general education undergraduate students and senior engineering students who volunteered to participate in the study. The participants completed three assessment instruments designed to measure engineering design performance. Two raters scored the assessments and the scores were tested for interrater reliability. These data were compared to existing data collected from middle school students as part of research on the EYE program. The data were analyzed for statistical, practical, and theoretical significance. A series of G theory analyses were used to evaluate the EYE assessment instruments' reliability in terms of overall scores and within each engineering dimension. Results are presented in the following chapter.

RESULTS

This chapter presents the results of the data analyses described in the methodology. Participant demographic and descriptive information is presented first followed by the results of the generalizability analyses and ANOVAs used to examine known-groups validity.

Demographics

Senior Engineering Students

The majority of participating engineering students were between ages 19 and 25; all of the students were male. Four participants identified with the age group 26-35 and two students were between age 36 and 45. Participants rated their interest in science/technology and engineering on a 5-point scale from one (not at all interested) to five (very interested). The average ratings for the engineering group were 4.5 for interest in science/technology and 4.5 for interest in engineering. Ten of the students participated in extra-curricular activities related to engineering in high school, six students took high school engineering classes, and the majority of the students took college engineering classes and participated in engineering extra-curricular activities during college.

General Education Undergraduate Students

All of the participating general education undergraduate students were between age 19-25 and nine of the participants were male. Participants rated their interest in science/technology and engineering on a 5-point scale from one (not at all interested) to five (very interested). The average scores for the general undergraduate group were 3.3 and 2.5, respectively. Two of the students participated in extra-curricular activities related to engineering in high school, two students had taken engineering classes in high school, and two had taken college-engineering classes and had participated in engineering extra-curricular activities during college. Six students in the general undergraduate group identified engineering as their major area of study.

Middle School Students

In addition, I included existing data from two groups of middle school students collected by Harlan et al. (2015) and Van Haneghan et al. (2015). This sample included 451 eighth grade students, 445 seventh grade students, and 422 sixth grade students. Approximately, half of the students were from schools participating in the EYE program and the remaining participants were from "control" schools not participating in the EYE program.

Generalizability and Reliability

Research question one was "Is there enough generalizability across people, independent of rater and form, to suggest an underlying general engineering design construct measured by the assessment instruments?"

The G coefficient indicates the reliability of the scores across raters and forms. The G coefficient for the reliability of total engineering performance scores and scores on the individual engineering dimensions are presented in Table 5.

Table 5

G Coefficients for Engineering Dimensions and Total Engineering Performance

Engineering Dimension	G Coefficient
Depth and Breadth of Thinking	.72
Teams and Expertise	.72
Evaluation of Design	.35
Use of Data and Research	.67
Total Engineering Performance	.89

Total scores on the engineering performance assessments yielded the largest G coefficient (ϕ = .89). This value indicates that the measurement instruments are reliable when compared to the conventional ϕ = .80 criterion for reliability (Mushquash & Conner, 2006).

G theory analyses were also conducted on scores from each dimension individually. Three of the dimensions had G coefficients close to ϕ =.7; all of the G coefficients were lower than the conventional ϕ = .80 criterion for reliability (depth and breadth of thinking, ϕ = .72; teams and expertise scores, ϕ = .72; critical evaluation of design, ϕ = .35; and use of data and research, ϕ = .67).

Table 6 shows that the $p \times f \times r$ design estimated seven variance components associated with the individual dimensions for engineering performance as well as variance components for total engineering performance.

Table 6

Variance Components for Engineering Performance Assessments

	Depth and Breadth of Thinking		Teams and Expertise		Evaluation of Design		Use of Data and Research		Total Engineering Performance	
Facet	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
p	.28	39	.27	37	.10	12	.34	36	4.55	52
f	.00	.00	.00	.00	.05	6	.05	5	.00	.00
r	.03	4	.06	7	.06	7	.02	2	.53	6
pf	.21	29	.24	33	.43	51	.39	42	.00	.00
pr	.03	4	.00	.00	.02	2	.03	3	.00	.00
fr	.01	1	.01	1	.00	0	.01	1	.32	4
pfr, e	.16	23	.16	22	.19	22	.11	11	3.38	39

For total engineering design performance, the largest estimated variance component was for person (σ^2 *p* = 4.55). This variance component is the estimated variation in respondents' scores when the score for each person represents his or her mean score across both raters and the three measurement forms. In other words, participants had differences in their engineering performance scores representing systematic individual differences in engineering performance. The next largest variance component for total engineering design scores, the three-way interaction (σ^2 *pfr* = 3.38), is a confounded and ambiguous term. It possibly indicates a three-way interaction between person, form, and rater. However, it also represents the residual and may be caused by facets that were not included in the analysis.

Results of four additional G analyses showed that person accounted for a moderate amount of variance, independent of rater and form, on the depth and breadth of thinking and teams and expertise dimensions (see Table 6). However, for the evaluation of design and use of data and research dimensions the majority of variance was accounted for by the person-form interaction indicating that different dimension performance scores may emerge, depending on which assessment form is used in one's analysis. The variance components for the three-way interactions (*pfr*) were also moderate on all dimensions. Again, this is a confounded and ambiguous term. It possibly indicates three-way interactions between person, form, and rater or may be caused by facets that were not included in the analysis. The variance components of rater and form accounted for vary little variance in each of the four engineering dimension G theory analyses.

Engineering Experience and Known Groups Validity

Engineering Experience and Total Engineering Design Performance

Three one-way ANOVAs (i.e., one for each engineering assessment) were used to compare assessment scores for the four different levels of engineering expertise: (1) middle school students enrolled in control schools, (2) middle school students enrolled in the EYE program, (3) general undergraduate students, and (4) senior engineering students. The results showed that data from all three ANOVAs did not violate the assumptions associated with one-way ANOVAs. Specifically, the standardized residuals for assessment scores were normally distributed and there was evidence supporting the homogeneity of variance assumption (i.e., the ratios of the largest to smallest group variances were less than 3:1). All mean comparisons were adjusted using the Bonferroni correction. The results of the ANOVAs are organized by assessment instrument and presented in the following sections.

Dog River assessment ANOVA. A one-way ANOVA was conducted to determine the effect of the four levels of engineering experience (control: *n* = 200, EYE: *n* = 222, general undergrad: *n* = 24, engineering: *n* = 23) on engineering design performance. The main effect of engineering experience was statistically significant, *F*(3, 468) = 30.25, *p* < .001, η^2 = .16.

As hypothesized, senior engineering students' scores (*M* = 10.57, *SD* = 1.80) were significantly higher than general undergraduate students (*M* = 7.13, *SD* = 2.35; *p* < .001) and middle school students (EYE: *M* = 6.04, *SD* = 2.71; control: *M* = 5.61, *SD* = 2.22; all *p* values < .001). General undergraduate students scored significantly higher than the control group of middle school students (*p* = .024). The difference between the middle school groups was not statistically significant (*p* > .05). Figure 1 illustrates the pattern of scores across groups.

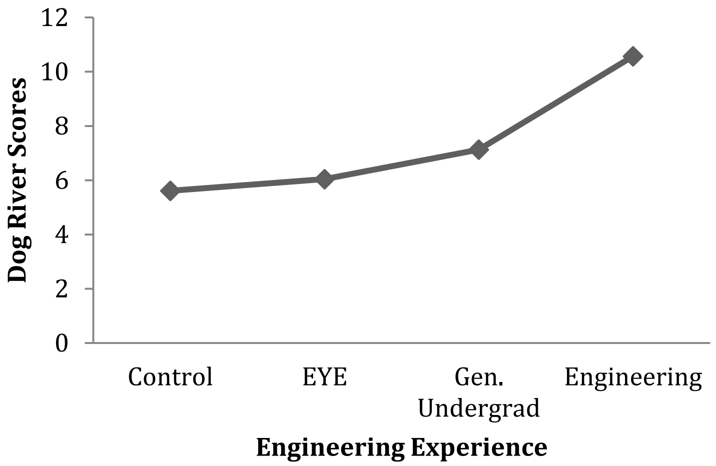


Figure 1. Group Mean Scores for the Dog River Engineering Assessment.

Seat belt assessment ANOVA. A one-way ANOVA was conducted to determine the effect of the four levels of engineering experience (control: *n* = 248, EYE: *n* = 203, general undergrad: *n* = 24, engineering: *n* = 23) on engineering design performance for the Seat Belt version of the engineering assessment. The main effect of engineering experience was statistically significant, *F*(3, 497) = 47.98, *p* < .001, η^2 = .23.

Post hoc tests were conducted to determine which differences between group means were statistically significant. Senior engineering students' scores (*M* = 8.98, *SD* = 2.25) were significantly higher than general undergraduate students (*M* = 7.08, *SD* = 2.60; *p* = .03) and general undergraduate students' scores were significantly higher than both middle school groups (control: *M* = 3.75, *SD* = 2.29; EYE: *M* = 4.24, *SD* = 2.28; all *p* values < .001). The difference between EYE students' scores the control group was not statistically significant. As hypothesized, the pattern of mean scores varied by experience level (see Figure 2).

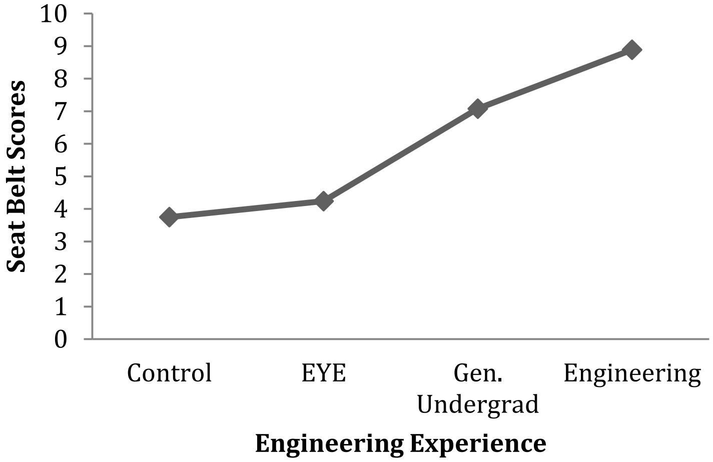


Figure 2. Group Mean Scores for the Seat Belt Engineering Assessment.

Algae assessment ANOVA. A one-way ANOVA was conducted to determine the effect of the four levels of engineering experience (control: n = 171, EYE: n = 274, general undergraduate: n = 24, engineering: n = 23) on engineering design performance for the Algae engineering assessment. The main effect of engineering experience was statistically significant, $F(3, 492) = 73.39$, $p < .001$, $\eta^2 = .31$.

Post hoc tests were conducted to determine which differences mean differences were statistically significant. Senior engineering students' scores (M = 10.43, SD = 1.80) were significantly higher than general undergraduate students (M = 8.17, SD = 1.95; p = .03) and general undergraduate students' scores were significantly higher than both middle school groups (control: M = 4.13, SD = 2.36; EYE: M = 4.54, SD = 2.20; all p values < .001). The difference between EYE students' scores and the control group was not statistically significant. As predicted, the pattern of mean scores varied by experience level (see Figure 3).

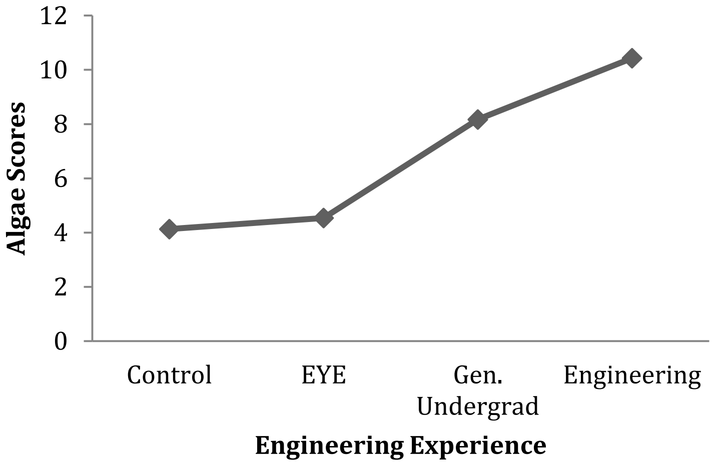


Figure 3. Group Mean Scores for the Algae Engineering Assessment

Engineering Experience and Engineering Dimension Performance

To further examine group differences, three 4 × 4 mixed ANOVAs were conducted with engineering experience as the between-subjects variable (four levels of engineering experience) and dimension as the within-subjects variable (four levels: depth and breadth of thinking, teams and expertise, critical evaluation of design, and use of data and research). A mixed ANOVA was conducted for each of the three assessments and the results are organized by assessment instrument.

Mauchley's test was conducted for each ANOVA and the results were statistically significant indicating the sphericity assumption for the within-subjects variable was violated in all three ANOVAs (Dog River: $\chi^2(5) = 26.00$, $p < .001$; Algae: $\chi^2(5) = 58.44$, $p < .001$; Seat Belt: $\chi^2(5) = 50.49$, $p < .001$). The results reported were adjusted using the Huynh-Feldt correction (Dog River: $\varepsilon = .94$; Algae: $\varepsilon = .98$; Seat Belt: $\varepsilon = .95$). Additionally, the normality assumption and the homogeneity of variance assumptions were examined and were not violated for the between-subjects variable in the ANOVAs.

Dog River assessment: Mixed ANOVA. A 4 × 4 mixed ANOVA was conducted to evaluate the effect of the four levels of engineering experience on scores for each of the four dimensions that contribute to the overall engineering assessment scores for the Dog River version of the engineering assessment. There were statistically significant main effects of group ($F(3, 465) = 30.25$, $p < .001$) and dimension ($F(2.81, 1307.58) = 18.54$, $p < .001$). The interaction between group and dimension was also statistically significant, $F(8.44, 1307.58) = 4.75$, $p < .001$. Therefore, the remainder of the results for this analysis will focus on the simple effects and paired comparisons between groups and within dimensions.

Dog River: Simple effect of group within each dimension. The simple effect of group was statistically significant within each of the four dimensions (all *p* values < .001) indicating that all of the groups did not perform equally on each of the individual dimensions. Mean scores on each dimension are presented in Table 7.

Table 7

Group Mean Scores for Dog River Dimensions

		Engineering Dimensions			
Group	*n*	Depth & Breadth	Teams & Expertise	Evaluation of Design	Data & Research
Control	222	1.79 (.63)	1.62 (.67)	1.21 (.94)	1.00 (.87)
EYE	200	1.87 (.84)	1.78 (.72)	1.16 (1.08)	1.24 (.94)
Gen. Undergrad	24	2.0 (.83)	1.67 (.87)	1.88 (.85)	2.74 (.54)
Engineering	23	2.61 (.78)	2.74 (.54)	2.74 (.54)	2.48 (.73)

Note. Standard deviations are in parentheses.

Paired comparisons determined which of the cell means were statistically different from the others. As seen in the Table 7, engineering students scored higher than all other experience groups on all four dimensions; the differences were statistically significant (all *p* values ≤ .03).

Dog River: Depth and breadth of thinking. Engineering students scored significantly higher than the other groups on the depth and breadth of thinking dimension (control: *p* < .001; EYE: *p* < .001; general undergraduate: *p* = .03). While the other groups followed the trend where control group scores were the lowest, the EYE scores were higher than the control group, and the general undergraduate scores were higher than the EYE scores, the differences between these groups were not statistically significant.

Dog River: Teams and expertise. Engineering students scored significantly higher than the other groups (all p values < .001). There were no statistically significant differences between the general undergraduate students, EYE students, and control group students (all p values > .05)

Dog River: Evaluation of design. Again, engineering students scored significantly higher than the other groups (control: $p < .001$; EYE: $p < .001$; general undergraduate: $p = .02$). General undergraduate students scored significantly higher than both the control ($p = .01$) and the EYE students ($p = .01$). However, there was not a statistically significant difference between the EYE and control group scores.

Dog River: Use of data and research. Again, engineering students scored significantly higher than the other groups (control: $p < .001$; EYE: $p < .001$; general undergraduate: $p = .004$). The control group scored significantly lower than the EYE ($p = .04$) and general undergraduate students ($p = .01$). The difference between general undergraduate scores and EYE scores was not statistically significant ($p > .05$).

Dog River: Simple effect of dimension within each group. Scores varied across dimensions for both of the middle school participant groups. The results showed that there were statistically significant simple effects of dimension for the control and EYE students (all p values < .001) indicating that the control and EYE groups did not have consistent scores across dimensions.

Dog River: Control group dimension scores. The control group's scores varied across all dimensions. Specifically, students in the control group scored higher on the depth and breadth of thinking dimension, followed by lower scores on the teams and expertise dimension, even lower scores on the critical evaluation of design dimension, and scored the lowest on the use of data and research dimension (all p values < .01).

Dog River: EYE group dimension scores. Students in the EYE group also scored differently across dimensions. They scored significantly higher on the depth and breadth of thinking and teams and expertise dimensions than on the evaluation of design and use of data and research dimensions (all p values < .001). However, they scored about the same on depth and breadth of thinking and teams and expertise (p > .05). The difference between data and research and evaluation of design were about the same (p > .05).

Dog River general undergraduate and engineering dimension scores. Interestingly, the more experienced groups had consistent scores across dimensions. For the general undergraduate and engineering students, there were no statistically significant differences in how they scored on the individual dimensions.

Seat belt assessment: Mixed ANOVA. A 4 × 4 mixed ANOVA was conducted to evaluate the effect of the four levels of engineering experience on scores for each of the four dimensions that contributed to the overall scores for the Seat Belt version of the engineering assessment. There were statistically significant main effects of group (F(3, 494) = 47.98, p < .001) and dimension (F(2.84, 1405.05) = 18.15, p < .001). The interaction between group and dimension was also statistically significant, F(8.53, 1405.05) = 3.65, p < .001. Therefore, the remainder of the results for this analysis will focus on simple effects and paired comparisons.

Seat Belt: Simple effect of group within each dimension. The simple effect of group was statistically significant within each of the four dimensions (all *p* values < .001) indicating that all groups did not perform equally on each of the individual dimensions. The mean scores for each group on each dimension are presented in Table 8.

Table 8

Group Mean Scores for Seat Belt Dimensions

		Engineering Dimensions			
Group	*n*	Depth & Breadth	Teams & Expertise	Evaluation of Design	Data & Research
Control	248	.96 (.85)	1.08 (.60)	1.21 (.95)	.50 (.76)
EYE	203	1.12 (.85)	1.04 (.59)	1.44 (.98)	.64 (.88)
Gen. Undergrad	24	1.80 (.88)	1.96 (.86)	1.67 (.87)	1.67 (1.13)
Engineering	23	2.22 (.85)	2.61 (.58)	2.22 (1.00)	1.91 (1.12)

Note. Standard deviations are in parentheses.

Paired comparisons determined which cell means were statistically different from the others. The results of the simple effects of experience level within each dimension are presented in the following paragraphs.

Seat belt: Depth and breadth of thinking. Engineering and general undergraduate students scored significantly higher than the two middle school groups (all *p* values < .003). The group means followed the trend where engineering students scored the highest followed by general undergraduate, EYE, and control group students, respectively. However, the difference between engineering and general undergraduate students' scores was not statistically significant. The difference between EYE and control group scores was also not statistically significant.

Seat belt: Teams and expertise. Scores within the teams and expertise dimension revealed statistically significant differences between groups. Engineering scores were significantly higher than the general undergraduate and control group scores (all *p* values < .01). General undergraduate students scored significantly higher than the control group as well (all *p* values < .003). The difference between control and EYE scores was not statistically significant.

Seat belt: Evaluation of design. While scores within this dimension followed the typical pattern of engineering students scoring the highest followed by general undergraduate, EYE, and control groups respectively, the only statistically significant differences were between engineering students and the two middle school groups (all *p* values < .002).

Seat belt: Use of data and research. Engineering and general undergraduate scores were about the same (*p* > .05) and both were significantly higher than the EYE and control group scores (all *p* values < .001). The two middle school groups also scored about the same (*p* > .05). EYE students scored slightly higher than the control group, but the difference was not statistically significant (*p* > .05).

Seat Belt: Simple effect of dimension within each group. Scores varied by dimension for some engineering experience groups. The results showed that there was a statistically significant simple effect of dimension for the engineering, EYE, and control groups (all *p* values < .003).

Seat belt: Engineering group dimension scores. The engineering group scores were mostly consistent across dimensions. The only statistically significant difference was that engineering students scored higher on the teams and expertise dimension than on the use of data and research dimension ($p = .001$).

Seat belt: EYE group dimension scores. Students in the EYE group also scored differently across dimensions. They scored significantly higher on the evaluation of design dimension than all other dimensions (all p values $< .001$). The EYE students scored about the same on depth and breadth of thinking and teams and expertise (all p values $> .05$) and the worst on the use of data and research dimension (all p values $< .001$).

Seat belt: Control group dimension scores. Similar to the EYE group, the control group scores were the lowest for the use of data and research dimension (all p values $< .001$). There were no significant differences between scores on depth and breadth of thinking, teams and expertise, and evaluation of design (all p values $> .05$).

Algae assessment: Mixed ANOVA. A 4 × 4 mixed ANOVA was conducted to evaluate the effect of the four levels of engineering experience on scores for each of the four dimensions that contribute to the overall engineering design performance for the algae version of the engineering assessment. There were statistically significant main effects of group ($F(3, 488) = 73.39$, $p < .001$) and dimension ($F(2.94, 1436.36) = 14.68$, $p < .001$). The interaction between group and dimension was also statistically significant, $F(8.83, 1436.3) = 7.76$, $p < .001$. Therefore, the remainder of the results for this analysis will focus on simple effects and paired comparisons.

Algae: Simple effect of group within each dimension. The simple effect of group was statistically significant within each of the four dimensions (all p values < .001) indicating that all of the groups did not perform equally on each of the individual dimensions. The group mean scores on each dimension are presented in Table 9. Paired comparisons were used to determine which of the cell means were statistically different from the others and the results are presented in the following paragraphs.

Table 9

Group Means for Dimensions on the Algae Assessment

		Engineering Dimensions			
Group	*n*	Depth & Breadth	Teams & Expertise	Evaluation of Design	Data & Research
Control	171	1.43 (.81)	1.23 (.73)	.76 (.69)	.72 (.98)
EYE	274	1.56 (.84)	1.48 (.65)	.79 (.77)	.72 (.98)
Gen. Undergrad	24	2.04 (.75)	1.96 (.91)	2.04 (.69)	2.13 (.68)
Engineering	23	2.48 (.79)	2.78 (.42)	2.52 (.73)	2.65 (.57)

Note. Standard deviations are in parentheses.

Algae: Depth and breadth of thinking. Engineering and general undergraduate students scored significantly higher than the two middle school groups (all p values < .003). The group means followed the trend where engineering students scored the highest followed by general undergraduate, EYE, and control group students, respectively. However, the difference between engineering and general undergraduate scores and the difference between EYE and control group scores were not statistically significant.

Algae: Teams and expertise. Scores within the teams and expertise dimension revealed statistically significant differences between all levels of experience. Engineering scores were the highest, followed by general undergraduate, EYE, and the control group scores (all p values < .003).

Algae: Evaluation of design. Engineering and general undergraduate students scored significantly higher than the two middle school groups (all p values < .001). The group means followed the trend where engineering students scored the highest followed by general undergraduate, EYE, and control group students, respectively. However, the differences between engineering and general undergraduate scores and between EYE and control group scores were not statistically significant.

Algae: Use of data and research. Engineering and general undergraduate scores were about the same (p > .05) and both were significantly higher than the EYE and control group scores (all p values < .001). EYE students scored slightly higher than control school students, but the difference was not statistically significant (p > .05).

Algae: Simple effect of dimension within each group. Scores varied across dimensions for both middle school participant groups. The results show that there were statistically significant simple effects of dimension for both the EYE and control groups (all p values < .001). The simple effects of dimension for the engineering and general undergraduate students were not statistically significant suggesting that students with more engineering experience tend to score about the same on all four dimensions.

Algae: EYE dimension scores. Students in the EYE group scored higher on the depth and breadth of thinking and teams and expertise dimensions than on the evaluation of design and use of data and research dimensions (all p values $< .001$). EYE students' scores were about the same for depth and breadth of thinking and teams and expertise ($p > .05$). They were also similar for evaluation of design and use of data and research ($p > .05$).

Algae: Control group dimension scores. Students in the control group also scored differently across dimensions. They scored the significantly higher on the depth and breadth of thinking dimension than teams and expertise ($p < .001$). They also scored significantly higher on teams and expertise than on both evaluation of design and use of data and research (all p values $< .001$). Scores on the evaluation of design and use of data and research dimensions were not significantly different ($p > .05$).

Chapter Summary

This chapter includes the results of a series of statistical analyses conducted during the current research. A G theory analysis on the data collected from the university students was used to evaluate the dependability or reliability of the assessment instruments. Overall, the reliability coefficient for the three assessment forms, with two raters, was higher than the conventional reliability criteria of $\phi = .80$. G theory analyses on the reliability of the individual dimensions were lower than the reliability criteria. Critical evaluation of an engineering design had the lowest G coefficient ($\phi = .35$).

A series of ANOVAs was conducted to evaluate group differences on engineering design performance overall and on engineering design performance across individual engineering dimensions. Group differences present across assessments and dimensions indicated that engineering experience does affect engineering design performance. However, group differences were inconsistent across instruments and dimensions. The results are summarized in the next chapter and presented with conclusions and recommendations for future research.

DISCUSSION

This chapter summarizes the findings from the previous chapter and presents the conclusions with recommendations for future research.

Summary of the Study

This study examined the relationship between engineering experience and engineering design performance and evaluated the psychometric properties of three assessment instruments designed to measure engineering design performance in association with the EYE program (Van Haneghan et al., 2015). The United States Congress Joint Economic Committee (2012) stated that technological skills are becoming more important to employers as technology becomes more integrated and more critical across industries. However, there is still a lack of available workers in STEM-related fields (Deloitte Consulting LLP, Oracle, & the Manufacturing Institute, 2009). In response to this deficit and a national shift toward increasing students' interest and performance in STEM fields, the Mobile County Public School system has begun to incorporate engineering instruction into existing middle school curricula.

The EYE modules were implemented by math and science teachers in two Mobile, Alabama middle schools to improve STEM performance and to increase students' interest in and beliefs about STEM fields and careers. The goal of the EYE program was to teach general engineering design skills, or “habits of mind,” that would transfer to novel design situations. Assessments play a critical role in all instructional systems because they determine if students have learned and are able to execute the performance objectives associated with the instruction (Pellegrino et al., 2014). Therefore, the EYE assessments aligned with the applied nature of the instructional goals to apply the engineering design process across content and outside of the classroom.

Harlan, Dean, et al. (2014) developed three assessment instruments to measure transfer of engineering design performance applied to relevant engineering design problems to assess the design process rather than the final design solution. The assessments were influenced by the works of Bailey and Szabo (2006) on evaluating design processes, Bransford and Schwartz’ (1999) theory of transfer, and Atman et al.’s (2007) findings related to information gathering and problem scoping.

While the EYE assessments do align with the engineering performance standards published in the NGSS (2013), it is important to note that the EYE program was developed prior to the publication of the NGSS. Clearly defined performance expectations are critical for the design and development of a comprehensive and cohesive instructional system and further research include an alignment effort with the formal science standards published in the NGSS (Pellegrino et al., 2014). Pellegrino et al. (2014) recommend some type of performance task to measure engineering design performance dimensions outlined in the NGSS. The EYE assessments appear to match the NGSS and assessment guidelines. However, more research would help to align the

EYE program instruction and assessments with the science and engineering standards outlined in the NGSS.

Initial research with middle school students on the reliability and validity of the EYE assessment instruments yielded positive results (Harlan, Dean, et al., 2014; Harlan, Pruet, et al., 2014; Harlan et al., 2015; Van Haneghan et al., 2015). However, there was a need to investigate the reliability and validity of the instruments. This study investigated the known-groups validity of the assessment instruments and the generalizability of scores on engineering performance with a group of college students with varied engineering experience.

Based on the published research findings presented in the literature review, I expected participants with more engineering experience to score higher on the engineering design assessments than students with less engineering experience (Ahmed, Wallace & Blessing, 2003; Atman & Bursic, 1996; Atman et al., 2007; Gruenther et al., 2009; Mullins et al., 1999). Generalizabilty of the assessment instruments was also examined in the current study and included multiple raters and forms as facets in the analysis. The college participants completed all three of the engineering assessments to allow for the G theory analysis of assessments. The G theory analyses were important because they allow for a more detailed analysis of measurement error and the sources of that error (Shavelson et al., 1989).

Conclusions drawn from the results of the G theory analyses and known-groups validity analyses are presented in the following sections with recommendations for future research. The goal of this study was to examine the validity and reliability of the EYE assessment instruments to provide evidence that sound interpretations can be drawn from the assessment results.

Discussion of the Generalizability Findings

A series of G theory analyses was used to evaluate the dependability, or reliability, of the assessments, which is especially important with applied assessments such as the engineering design assessments (Briesch et al., 2014). The G coefficient for the three assessment forms with two raters was higher than the conventional .80 criterion for reliability indices (Shavelson et al., 1989). The G theory analyses also revealed that rater and form did not account for large portions of score variation; this provided further evidence of the reliability of the instruments suggesting that results from the assessment instruments generalize to the general construct of engineering design performance.

Individual dimension G theory analyses revealed some reliability deficiencies within the individual dimensions. Of particular concern, the evaluation of design dimension had the lowest G coefficient. The remaining three dimensions generated G coefficients that were only slightly below the conventional reliability criterion. Because the results for the individual dimensions were less reliable, I recommend that the questions and scoring rubric be revised to potentially increase reliability and generalizability across dimensions and possibly improve the overall reliability of the instruments. In particular, revising the evaluation of design questions to better capture the construct and increasing rubric clarity for that dimension could increase reliability.

Discussion of the Known-Groups Validity Findings

Engineering Experience and Total Engineering Design Performance

The results of the known-groups validity analyses for total engineering design performance were consistent with prior research and suggest that engineering experience affects engineering design performance. Total group scores followed the expected pattern on all three instruments. As hypothesized, it was determined that as engineering experience increases, engineering design performance tends to improve as well. When considering total engineering scores, the engineering students scored higher than all other groups on all three assessments.

Although the EYE middle school students scored slightly higher than students did in the control middle schools on all three of the assessments, the differences were not statistically significant. However, in a study conducted by Van Haneghan et al. (2015), significant differences existed between the middle school students only on certain engineering design dimensions. Specifically, students who had participated in the EYE modules performed significantly better than the control group on three engineering design dimensions: depth and breadth of thinking, critical evaluation of the design, and use of data and research.

The findings from the known-groups validity analyses add support to the literature showing that engineering design experience improves engineering design performance (e.g., Atman & Bursic, 1996; Gruenther et al, 2009; Mullins et al., 1999) and are consistent with the previous research conducted by Van Haneghan et al. (2015). In addition, the results of these ANOVAs add support to the validity of the assessment instruments used in association with the EYE program. The assessment instruments were expected to differentiate engineering performance by level of engineering experience. While the instruments did not identify differences between middle school students' scores, they did recognize group differences between more extreme engineering experience variations, thus providing weak to moderate evidence of known-groups validity.

There were limitations to the group characteristics that could have influenced the results. First, the two middle school samples were originally matched to be as equivalent groups as possible. At the same time the EYE was incorporated in participating schools, the Mobile County school district reformed its middle school curriculum standards to include engineering design instruction for all students, including the students at the control schools (Harlan et al., 2015). Therefore, the control schools were not a true matched comparison group as originally designed. This likely influenced the limited number of statistically significant differences between the EYE and control group engineering design scores.

A better design would have been to either randomly sample students from multiple schools with and without the EYE program or to conduct pretests and posttests before and after students completed an EYE module. Harlan et al. (2015) proposed to study the modules as a randomized clinical trial across middle schools. This would reduce the impact extraneous factors (e.g., teachers, school environment) on the results and provide more valid and generalizable results.

Second, there were some general undergraduate students with engineering experience, a few engineering students who did not show a strong interest in engineering, a few engineering students who reported less engineering experience than expected, and a limited scope of engineering students (i.e., only civil engineering students). Developing stricter criteria for group membership would help to minimize within group differences in engineering experience and potentially improve the validity of the assessment data. Also, Harlan et al. (2015) and Van Haneghan et al. (2015) controlled for math and reading ability using students' fifth grade math and reading standardized test scores. The current study did not attempt to control for individual differences in math and reading, which likely affected the results. In the future, I recommend including college students' ACT scores as a control variable to control for individual differences in English, math, reading, and science ability.

The results of the current research provide a glimpse into engineering design performance differences across groups. Because the study included a limited and targeted convenience sample it is recommended that a similar study be conducted, building on these results, with better-defined groups and with additional groups.

Discussion of Engineering Experience and Engineering Dimension Results

In addition to overall engineering design performance, group performance differences were examined within each engineering design dimension (depth and breadth of thinking, teams and expertise, critical evaluation of design, and use of data and research) across the three engineering assessments (Dog River, Seat Belt, and Algae assessment instruments used with the EYE program). Perhaps the most notable finding was that engineering students tended to score higher than the other groups and had consistently high scores within each individual dimension. General undergraduate students also had fairly consistent scores across the individual dimensions. The less experienced middle school students, however, had more variation in how they performed on the different dimensions.

Data from all three assessments showed that the middle school students performed the worst when asked to identify ways to use data and research and when asked to identify data that would help them develop a solution. They consistently scored higher on depth and breadth of thinking and teams and expertise dimensions than on the use of data and research dimension. Scores on the evaluation of a design dimension varied more depending on the individual assessment scenario. For example, the middle school students tended to score about the same on depth and breadth of thinking, teams and expertise, and evaluation of design for the algae scenario, but scored lower on evaluation of design when answering the Dog River scenario questions.

Group differences were present on all four dimensions on all three assessments. The general pattern of dimension scores was that engineering students scored the highest, followed by general undergraduate students, EYE students, and control group students. However, there were variations in which differences were statistically significant across dimensions and assessment scenarios. For example, the EYE students scored better than the control group students only on the teams and expertise dimensions on the algae assessment, the evaluation of design on the seat belt scenario, and on evaluation of design and use of data and research dimension on the Dog River assessment.

Van Haneghan et al. (2015) found that EYE participants scored higher on three engineering design dimensions: depth and breadth of thinking, critical evaluation of the design, and use of data and research. These results were replicated in the current study only when paired with particular scenarios. On the majority of dimensions across scenarios, there were few statistically significant differences between the EYE students and control group students.

The differences between general undergraduate students' scores and engineering students' scores also varied according to assessment scenario. When using the Dog River assessment scenario, engineering students scored higher than general undergraduate students did on all dimensions. The engineering students who participated in the study were civil engineering students, which could have affected the scenario-based variability as the Dog River assessment fits best with the civil engineering field. However, when using the seat belt and algae assessment scenarios, engineering students scored about the same as general undergraduate students on all dimensions except that they scored higher on the teams and expertise dimension. These inconsistencies could also be a result of ill-defined group characteristics. There were several general undergraduate students who listed engineering as their major field of study and showed high levels of interest in both engineering and math.

Recommendations for Future Research

This study focused on the generalizability and the validity of three engineering design assessment instruments developed for the EYE program. Taken together as a group of three assessment forms measuring overall engineering design performance, the instruments can be considered reliable instruments. However, the results of the analyses were inconsistent across groups and dimensions, and the individual dimension reliability coefficients were low. There was weak to moderate evidence of known-groups validity because the group differences varied across assessment forms and dimensions.

These assessment instruments require more research on how they would function within various populations. Also, the questions and scoring criteria need to be revised and retested in an effort to increase reliability. Only weak to moderate evidence of known-groups validity was found. Researching these assessment instruments with more definitive group parameters potentially could enhance the validity of the instruments. Recommendations for future research are presented in more detail in the following sections.

Generalizability and Reliability Research

The results of the generalizability analyses suggest that the engineering design assessments, as a whole, are reliable. However, the individual engineering dimensions revealed G coefficients that were lower than the conventional criterion for reliability. In the future, the questions and/or how the questions are scored should be revised to better capture the dimensional constructs and should be retested for potentially increased reliability indices. Revising the rubric could potentially increase interrater reliability as well. Harlan, Dean, et al. (2014) found moderate to substantial interrater reliability; this was not replicated in the current study. More clearly defined scoring criteria in the rubric could result in higher and more consistent interrater reliability of the instrument.

The G theory analyses included data from the college students only because the middle school students did not complete all three of the assessments. Therefore, there is no evidence of assessment instrument reliability if used with other populations. Specifically, the instruments were designed for administration to middle school students in conjunction with the EYE program. Generalizability of the instruments was not determinable with the data set provided by Van Haneghan et al. (2015). In the future, I recommend that middle school students complete all three assessments in order to evaluate the reliability of the instruments within the target population. It would also be valuable to administer the assessments to other populations with varying levels of engineering experience and different age groups and continue to test, revise, and retest the assessment to improve construct validity and reliability.

Validity Research

Several factors may have influenced the weak to moderate evidence of known-groups validity. Most importantly, the definitive group difference between the EYE and control groups was compromised when the control schools started incorporating STEM instruction into their curricula and there were few statistically significant differences between the middle school groups. There were also similarities between the general undergraduate and engineering students' performance on several dimensions with most differences occurring when using the Dog River assessment. The Dog River assessment most closely aligned with the civil engineering focus of study within the engineering student group. This likely explains the more pronounced difference between engineering and general undergraduate students when using the Dog River assessment.

Similarities between the engineering and general undergraduate students on the seatbelt and algae assessments may have been a result convenience sampling, as there were several engineering students in the general undergraduate group and the sample of engineering students was limited to only civil engineering students. In addition, there were several students in the engineering group that answered with lower ratings for interest in science and engineering than expected. In the future, larger sample sizes and more restrictive group criteria could possibly change the known-groups validity evidence for the assessments.

Although there were inconsistencies in groups' performance across assessment instruments and dimensions, the results of the current study do provide some evidence of known-groups validity. The instruments differentiated performance differences between engineering students and other groups when considering total engineering design performance scores. Results were less consistent when considering individual dimensions. The overall scores from two raters on three instruments provided evidence of reliability of the instruments. Again, the results were less consistent when evaluating the reliability of the individual dimensions.

A summary of the recommendations for future research follows:

1. Revise questions to measure the dimensional constructs more accurately and to increase reliability.
2. Revise scoring rubric to increase interrater reliability.
3. Administer all three assessments to middle school students to allow for a generalizability analysis with the target audience.
4. Administer the assessments to other populations with varying levels of engineering experience and different age groups.
5. Use random sampling and larger sample sizes.
6. Use more conservative and explicit group characteristics to better define engineering experience levels.
7. Administer pretests and posttests to EYE participants for a more powerful research design.

Implications for Instructional Design

Design is a central component of engineering, but it can be hard to teach and difficult to develop valid and reliable assessments (Cardella et al., 2011; Dym et al., 2005). Instruction should support the performance objectives and learning assessments should provide evidence of learning and proficiency. During assessment design, it is critical for instructional designers to begin with clearly defined performance expectations that represent how the learner will use the knowledge or skills in the real world to facilitate retention and transfer (Gagne, 1972 as cited in Reiser & Dempsey, 2007; Pellegrino et al., 2014). For example, the EYE program utilized repetitious exposure and practice with the engineering design process using relevant application opportunities to enhance retention and transfer of the four engineering habits of mind to novel situations. The EYE assessments measured engineering design performance aligned with the performance expectations.

However, it is important for instructional designers to know that no matter how carefully assessments are designed to align with the performance expectations, it is imperative to investigate the validity and reliability of the assessments (Messick, 1989). This is a necessary step to ensure that the interpretations and uses of assessment results accurately represent the skills measured by the assessments and to develop a comprehensive instructional system.

The purpose of this study was to investigate the validity and reliability of three engineering assessment instruments that were carefully designed to measure knowledge and skills associated with the engineering design performance expectations. While I did find some evidence of validity and reliability associated with the instruments, I also found that more work is required to better capture the constructs of engineering design and the associated dimensions. The results illustrate the importance of validity and reliability analyses to isolate systematic error associated with assessment instruments and to evaluate how well the instruments measure the target constructs.

Conclusion

As the world becomes increasingly technological, the United States must focus on developing a STEM-competent workforce to gain and maintain a competitive advantage in the global market place. Cultivating STEM-competent workers, innovators, problem-solvers, and critical thinkers is necessary to solve incredible challenges such as energy, health, environmental protection, and national security (PCAST, 2010) and this depends on the effectiveness of STEM education.

The United States has lagged behind other nations in STEM fields, interest in STEM, and STEM proficiency but is taking action to improve the STEM workforce through better STEM education. The current study provides an extension of research conducted in association with the Engaging Youth through Engineering program incorporated into select schools in Mobile, Alabama to increase interest and to improve performance in engineering. A critical step to develop and implement comprehensive engineering programs is to measure performance against standards using valid and reliable assessments because, without valid and reliable assessments there is no way to know that STEM education programs are increasing interest or proficiency in STEM. Valid and reliable assessment instruments allow researchers and consumers to trust the conclusions drawn from assessment results and properly evaluate the success of the instruction.

The findings from the current study inform how to administer the EYE engineering assessments to provide valid and reliable measurement of engineering design performance. The results suggest that using the three assessments together and considering overall engineering design performance provides a valid and reliable measure of engineering design performance. This study also identified ways to potentially improve the reliability and validity of the assessment instruments.

The EYE assessments, and associated instruction, are in the early stages of implementation and testing. It is important that the assessments are revised and retested to potentially improve the quality of the inferences drawn about engineering design performance and accurately assess the success of the EYE program. This study and the associated findings are a small but important step toward improved interest and proficiency in STEM and are a step in the right direction toward the U.S. gaining a more competitive position in the global market place.

REFERENCES

REFERENCES

Accreditation Board for Engineering and Technology. (2000). *Criteria for Accrediting Engineering Programs.* Retrieved from http://www.abet.org

Ahmed, S., Wallace, K. M., & Blessing, L. T. M. (2003). Understanding the differences between how novice and experienced designers approach design tasks. *Research in Engineering Design, 14*, 1-11. doi:10.1007/s00163-002-0023-z

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

Archibald, D. A., & Newman, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in secondary schools*. Washington, DC: National Association of Secondary School Principals.

Atman, C. J., Adams, R. S., Cardella, M. E., Turns, J., Mosborg, S., & Saleem, J. (2007). Engineering design processes: A comparison of students and expert practitioners. *Journal of Engineering Education, 96*, 359-379. doi:10.1002/j.2168-9830.2007.tb00945.x

Atman, C. J. & Bursic, K. M. (1996). Teaching engineering design: Can reading a textbook make a difference? *Research in Engineering Design, 8*, 240-250.

Atman, C. J., Cardella, M. E., Turns, J., & Adams, R. (2005). Comparing freshman and senior engineering design processes: An in-depth follow-up study. *Design Studies, 26*, 325-357. doi:10.1016/j.destud.2004.09.005

Atman, C. J., Chimka, J. R., Bursic, K. M., & Nachtmann, H. L. (1999). A comparison of freshman and senior engineering design processes. *Design Studies, 20*, 131-152. doi:10.1016/s0142-694x(98)00031-3

Bailey, R. (2008). Comparative study of undergraduate and practicing engineer knowledge of the roles of problem definition and idea generation in design. *International Journal of Engineering Education*, *24*, 226-233.

Bailey, R., & Szabo, Z. (2006). Assessing engineering design process knowledge. *International Journal of Engineering Education, 22*, 508-518. Retrieved from http://www.ijee.dit.ie/

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: brain, mind, experience, and school*. Washington, D.C.: National Academy Press. doi:10.17226/9853

Bransford, J. D. & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Educaion, 24*, 61-100. doi:10.2307/1167267

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*, 13-35. doi:10.1016/j.jsp.2013.11.008

Broudy, H. S. (1977). Types of knowledge and purposes of education. In R. C. Anderson, R. J. Spiro, & W. E. Montegue (Eds.), Schooling and the acquisition of knowledge. Hillsdale, NJ: Erlbaum.

Cardella, M. E., Oakes, W. C., Zoltowski, C. B., Adams, R., Purzer, S., Borgford-Parnell, J., Bailey, R., & Davis, D. (2011, October). *Assessing student learning of engineering design.* Special session at the 41st meeting of ASEE/IEEE Frontiers in Education Conference, Rapid City, SD. doi:10.1109/fie.2011.6143109

Commission on Mathematics and Science Education (2010). *The opportunity equation: Transforming mathematics and science education for citizenship and the global economy*. Carnegie Corporation of New York. Retrieved from https://www.carnegie.org

Cook, D. A. (2014). Much ado about differences: Why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education, 20*, 829-834. doi:10.1007/s10459-014-9551-3

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine, 88*, 872-883. doi:10.1097/acm.0b013e31828ffdcf

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302. doi:10.1037/h0040957

Deloitte Consulting LLP, Oracle, & the Manufacturing Institute. (2009). *People and profitability: A time for change*. Retrieved from https://www2.deloitte.com

Dick, W., Carey. L., & Carey, J. O. (2009). *The systematic design of instruction.* Upper Saddle River, N. J.: Merrill/Pearson.

Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., & Leifer, L. J (2005). Engineering design thinking, teaching, and learning. *Journal of Engineering Education*, 103-120. doi:10.1002/j.2168-9830.2005.tb00832.x

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research, 3,* 430-441. doi:10.1177/104973239300300403

Fricke, G. (1999). Successful approaches in dealing with differently precise design problems. *Design Studies, 20*, 417-429. doi:10.1016/s0142-694x(99)00018-6

Gruenther, K., Bailey, R., Wilson, J., Plucker, C., and Hashmi, H. (2009). The influence of prior industry experience and multidisciplinary teamwork on student design learning in a capstone design course. *Design Studies*, *30*, 721-736. doi:10.1016/j.destud.2009.06.001

Harlan, J. M., Dean, M., Van Haneghan, J. P. (2014). Development of a rubric for use in assessing transfer of learning in middle grades engineering program participants. *Proceedings of the 2014 American Society for Engineering Education Gulf-Southwest Conference, Tulane University, New Orleans, LA*. Retrieved from http://asee-gsw.tulane.edu

Harlan, J. M., Pruet, S. A., Van Haneghan, J. P., and Dean, M. D. (2014, June). Using curriculum-integrated engineering modules to improve understanding of math and science content and STEM attitudes in middle grade students. *Proceedings of the*

2014 American Society for Engineering Education Gulf-Southwest Conference, Tulane University, New Orleans, LA. Retrieved from http://asee-gsw.tulane.edu

Harlan, J. M., Van Haneghan, J., Dean, M. D., & Pruet, S. A. (2015, June). *Evaluating the Impact of Curriculum-Integrated Engineering Design Modules in Middle Grades Classrooms*. Paper presented at 2015 ASEE Annual Conference & Exposition, Seattle, WA. doi: 10.18260/p.24028

Hattie, J. & Cooksey, R. W. (1984). Procedures for assessing the validities of test using the "known-groups" method. *Applied Psychological measurement, 8,* 295-305. doi:10.1177/014662168400800306

Hibberts, M., Johnson, R. B., & Hudson, K. (2012). Common survey sampling techniques. In *Handbood of survey methodology for the social sciences* (pp. 53-74). New York, NY: Springer. doi:10.1007/978-1-4614-3876-2_5

Honey, M., Pearson, G., & Schweingruber, H. (Eds.) (2014). *STEM Integration in K-12 Education: Status, Prospects, and an Agenda for Research*. Washington, DC: The National Academies Press. doi:10.17226/18612

International Technology and Engineering Educators Association. (2006). *Engineering by design: A guide to a national standards-based program model*. Reston, VA. Retrieved from https://www.iteea.org/STEMCenter/EbD.aspx

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73. doi:10.1111/jedm.12000

Katehi, L., Pearson, G., & Feder, M. (2009). *Engineering in K-12 education*. Washington, DC: The National Academies Press. doi: 10.17226/12635

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*, 15-23. doi:10.2307/1176232

Mayer, R. E. (1999). Designing instruction for constructivist learning. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (141-159). New York, NY: Routledge.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11. doi:10.2307/1175249

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*, 13-23. doi:10.2307/1176219

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: 10.1002/j.2333-8504.1994.tb01618.x

Michael, A. L., Klee, T., Bransford, J. D., & Warren, S. (1993). The transition from theory to therapy: Test of instructional methods. *Applied Cognitive Psychology, 7*, 139-154. doi:10.1002/acp.2350070206

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement, 53*, 265-292. doi:10.1111/jedm.12117

Mullins, C. A., Atman, C. J., & Shuman, L. J. (1999). Freshman engineers' performance when solving design problems. *IEEE Transactions on Education, 42*, 281-288. doi:10.1109/13.804533

Museum of Science. (2005). *Engineering is Elementary*. Boston, MA. www.mos.org/eie

Mushquash, C. & O'Connor, B (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*, 542-547. doi:10.3758/bf03192810

National Academy of Engineering. (2008). *NAE Grand Challenges for Engineering.* Retrieved from http://www.engineeringchallenges.ord/cms/challenges.aspx

National Academy of Science. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academies Press. doi:10.17226/11463

National Center for Education Statistics (2014). *National assessment of educational progress: The nation's report card*. Washington, DC: Institute of Educational Sciences, Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states* (vol. 1, The Standards). Washington, DC: The National Academies Press. doi:10.17226/18290

Organisation for Economic Co-operation, Development (OECD). (2008). Encouraging student interest in science and technology studies. Paris, France: Author. doi:10.1787/9789264040892-en

Pellegrino, J. M., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, D.C.: The National Academies Press. doi:10.17226/18409

Pinnell, M., Rowly, J., Preiss, S., Franco, S., Blust, R., & Beach, R. (2013). Bridging the gap between engineering design and PK-12 curriculum development through the use of STEM education quality framework. *Journal of STEM Education, 14,* 28-35. Retrieved from http://ojs.jstem.org/index.php?journal=JSTEM&page=article&op=view&path%5B%5D=1804&path%5B%5D=1562

President's Council of Advisors on Science and Technology (PCAST). September, 2010. Prepare and inspire: K-12 Science, Technology, Engineering and Math (STEM) education for America's Future. Retrieved from www.whitehousegov/ostp/pcast

Project Lead the Way. (2005*). About project lead the way: An overview*. Clifton Park, NY. Retrieved from https://www.pltw.org/our-programs/pltw-engineering

Puente, S. M., van Eijck, M., & Jochems, W. (2013). A sampled literature review of design-based learning approaches: A search for key characteristics. *International Journal of Technology and Design Education, 23,* 717-732. doi:10.1007/s10798-012-9212-x

Razzouk, R. & Shute, V. (2012). What is design thinking and why is it important? *Review of Educational Research, 82,* 330-348. doi:10.3102/0034654312457429

Reiser, R. A., & Dempsey, J. V. (Eds.). (2007). Trends and issues in instructional design and technology. Upper Saddle River, NJ: Pearson Education, Inc.

Rennie, L., Venville, G., & Wallace, J. (Eds.). (2012). Integrating science technology engineering and mathematics: Issues, reflections, and ways forward. New York, NY: Routledge. doi:10.4324/9780203803899

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Los Angeles: Sage. doi:10.7748/nr.12.4.86.s2

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922. doi:10.1037/0003-066x.44.6.922

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20,* 2-16. doi:10.2307/1177000

U.S. Congress Joint Economic Committee. (2012). *STEM education: Preparing for the jobs of the future*. Washington, DC. Retrieved from https://www.jec.senate.gov/public/index.cfm/democrats/2012/4/stem-education-preparing-jobs-of-the-future

Van Haneghan, J. P., Harlan, J. M., & Dean, M. D. (2015, April). *The impact of engineering focused modules on the engineering design knowledge of eighth graders.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Vogt, W. P. & Johnson, R. B. (2011). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences.* Los Angeles: Sage. doi:10.5860/choice.43-0697

Wolf, D. P. (1992). Good measure: Assessment as a tool for educational reform. *Educational Leadership, 49,* 8-13. Retrieved from Academic Search Complete, EBSCOhost

Zaidi, N. L. B., Swoboda, C. M, Kelcey, B. M., & Manuel, R. S. (2017). Hidden item variance in multiple mini-interview scores. *Advances in Health Sciences Education, 22,* 337-363. doi:10.1007/s10459-016-9706-5

APPENDICES

Appendix A – Engineering Design Assessment Instruments

ALGAE TASK

There has been a great deal of interest in developing alternative sources of energy. One idea is to use plants or other organisms to create biofuel. Theresa and Thomas want to produce biofuels. They hear from a friend that some engineers think algae might be a good source of biofuel. They want to design a way to grow enough algae in one year to produce fuel to fill a 10,000 gallon tank.

1. What kind of information do you think would be helpful for Theresa and Thomas to know before they get started to help them solve this problem?

   What questions should they ask before beginning to find a solution?

2. Theresa and Thomas want to put together a team to work on solving this problem. What kinds of experts would need to be on the team?

   What kinds of knowledge and skills would their team need to have?

3. Here are the steps Theresa and Thomas took to solve this problem:

   - They bought a pond in a rural area of South Alabama.
   - They grew algae in the pond.
   - They used the algae to make biofuel.
   - They measured how much biofuel they were able to make.

They found out that they did not make enough biofuel to meet their goal. Now, Theresa and Thomas have asked you to help them figure out where they went wrong and help them meet their goals.

What did they do well?

What steps, if any, were missing from their attempt to solve this problem?

Thomas and Theresa agree that they want to improve their design to solve the problem: grow enough algae in one year to produce 10,000 gallons of biofuel. What are some ways they could improve their process?

4. The two graphs and the table on this page are from some research on using algae for biofuel.

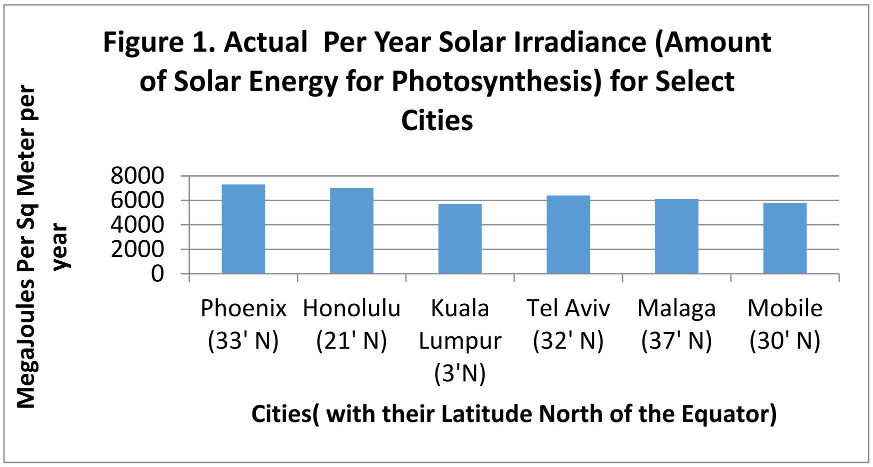


Figure 1. Actual Per Year Solar Irradiance (Amount of Solar Energy for Photosynthesis) for Select Cities

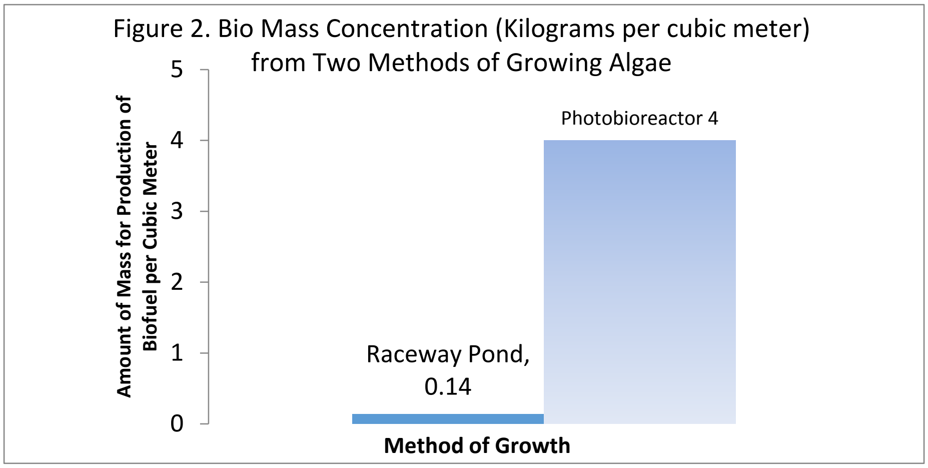


Figure 2. Bio Mass Concentration (Kilograms per cubic meter) from Two Methods of Growing Algae

Table 1. % Dry Weight Oil Content (for making Fuel) of Microalgae by Species of Algae

(% dry weight for making biofuel—
Range of values from studies)

Species of algae	**Lowest Value found**	**Highest Value Found**
Species 1	25%	75%
Species 2	28%	32%
Species 3	35%	54%
Species 4	45%	47%
Species 5	50%	77%
Species 6	15%	23%

How could you use the information in these graphs to help you develop a better solution to the problem?

What other data do you think might help you solve the problem? What else might you need to do research on to solve this problem?

DOG RIVER TRASH TASK

Dog River in Mobile, AL is typically littered with trash after a heavy rainfall. The city, the county, and the state want to know how to solve this problem, and have asked Charles, a local business owner, to help create a solution to this problem.

1. What information do you think would be helpful for Charles to know before he gets started to help him solve this problem?

   What questions does he need to ask?

2. Charles wants to put together a team to work on solving this problem. What kinds of experts would need to be on the team?

   What kinds of knowledge and skills would their team need to have?

3. Below is a description of the steps the team used to come up with a solution to this problem.

Step 1: They started off by discussing the problem. They decided that it was caused by litter from streams that fed into the Dog River.

Step 2: One team member designed a machine that could be pulled behind a boat that would “rake up” the litter.

Step 3: They built three of the machines.

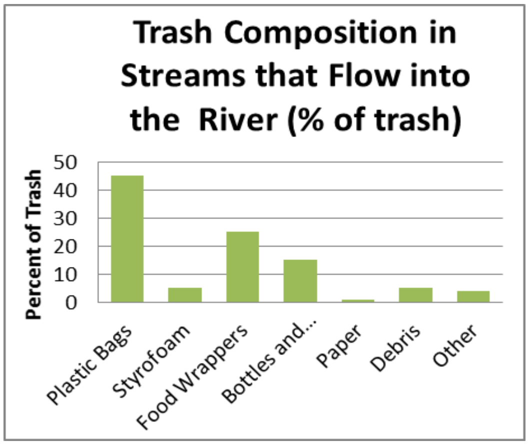
Step 4: They hired three boaters to pull the machines behind their boats.

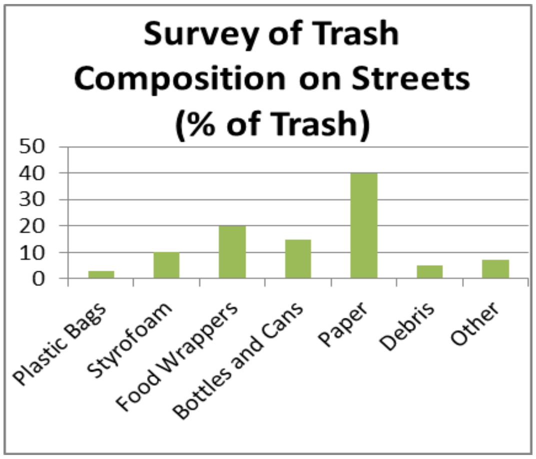
What did they do well?

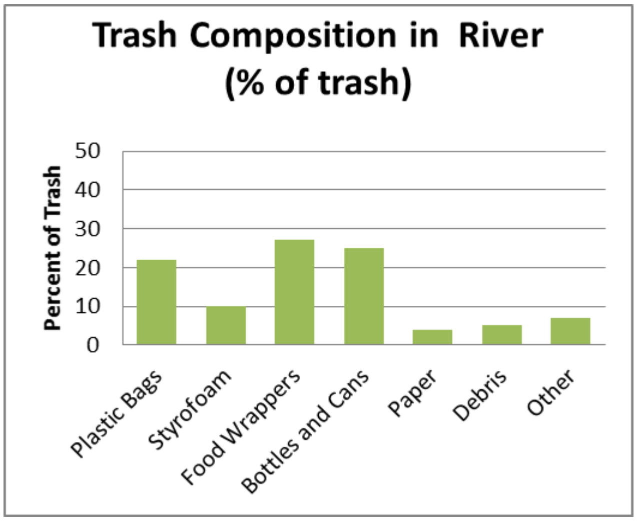
What steps, if any, were missing from their attempt to solve this problem?

What are some ways they could improve their process?

4. The three graphs on this page are from a study of river trash







How could you use the information in these graphs to help you develop a better solution to the problem?

What other data do you think might help you solve the problem? What else might you need to do research on to solve this problem?

SEAT BELT TASK

Julie learns that her grandmother has broken several ribs in an automobile accident. Julie is puzzled by the injuries, because the accident happened at a low speed (20 miles per hour) and her grandmother was wearing her seatbelt. When Julie talked to the doctors, they told her that the injury appeared to be caused by the seatbelt, and that they often see this type of injury in elderly people who have been in an accident. Julie thinks to herself "Somebody should invent a better seatbelt that reduces the risk of these injuries in elderly people." She decides that when she goes to work tomorrow morning, she is going to put together a team to solve this problem.

1. What information do you think would be helpful for Julie to know before she gets started to help her solve this problem?

   What questions does she need to ask?

2. Julie wants to put together a team to work on solving this problem. What kinds of experts would need to be on the team?

   What kinds of knowledge and skills would their team need to have?

3. The team gets together and takes the following steps:

   a. They discuss what the problem is.
   b. They decide to focus on finding a way to adapt the vehicles that people already have.
   c. Next, they discuss how they will do it. Kristin, who is part of the team, remembers that she got a package last week, and it had foam inside to keep things inside from breaking.
   d. They decide to make foam cushions that can be fit onto existing shoulder and lap belts.
   e. The team develops a set of foam cushions that can be put onto currently existing lap and shoulder belts.
   f. After developing the cushions, they look for someone to help sell them.

   What did they do well?
   What steps, if any, were missing from their attempt to solve this problem?

What are some ways they could improve their process?

4. The graphs and table on this page are from research about age, force of car accidents, and restraint types.

Figure 1. Force of seatbelt, force of airbag, and chest deflection with different types of restraints

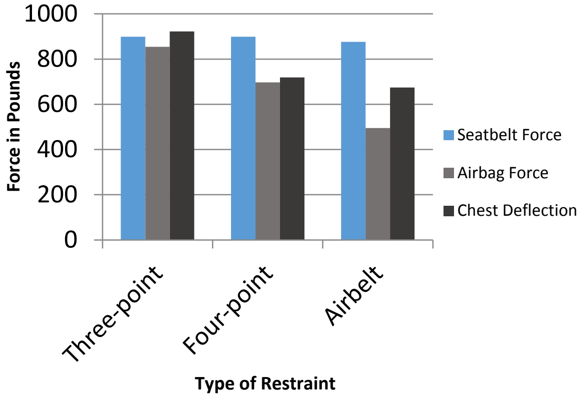
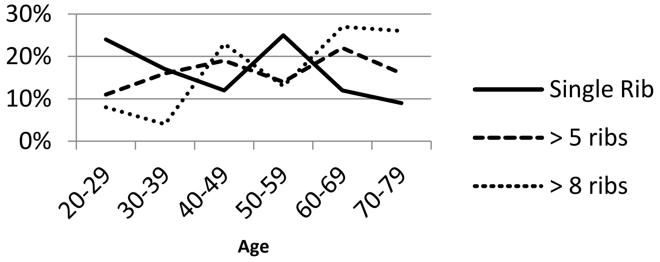


Figure 2. Age Distribution of Rib Fractures in Frontal Crash Occupants aged 20-79



Type of Material	Density	Level of Deformity Under Force	Compression
Polyurethane foam	25.6	1600	0.44
IMPAXX foam	33.7	3400	0.40
Aluminum foam	470	69000	0.29
Cork	293	9000	0.30

How could you use the information to help you figure out a better solution to the problem?

What other data do you think might help you solve the problem? What else might you need to do research on to solve this problem?

Appendix B – Sample Scoring Rubric (Seat Belt Task)

The problem: Elderly people are injured in car accidents by their seatbelts. Julie wants her team to invent a better seatbelt to reduce the risk of these injuries in elderly people.

<table>
<tr><th colspan="4">Dimension 1: Depth and Breadth of Thinking</th></tr>
<tr><td colspan="4">What does she need to know or ask about the problem before getting started? Generally, you should use the student's response to Question 1 to score their depth and breadth of understanding about the problem. However, when you score this item, consider the student's responses in the context of their overall responses on the assessment. For example, in some cases, the student may reference the speed of the grandmother's car. This statement, in and of itself, may indicate that the student is focusing on what happened to the grandmother (past) and is not identifying information needed or questions to ask in order to find a solution. If this is the case, the student should be given a score of "0". However, the student might connect the issue of vehicle speed to the type of experts needed, development of a solution, or data/research needed. In this case, the student's score should reflect the number and integration of the aspects mentioned by the students.</td></tr>
<tr><td>0</td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>No answer or irrelevant responses. (e.g., response is not related to finding a solution or focuses on the past rather than on finding a solution).</td><td>Mentions something specific about the problem or information needed to solve it, but the response only examines one aspect without consideration of the bigger problem. For example, the student may suggest that Julie should find out what type of seatbelts cause these injuries.</td><td>Mentions multiple aspects of the problem or information needed to find a solution but does not integrate them in to a systematic approach to addressing the problem. For example, the student may identify the need to understand how frequently these injuries happen and the typical type of vehicle in which the injuries occur, but fails to address issues related to passengers, deign of the seatbelt, or manufacturing or marketing issues.</td><td>Response shows an integrated view of the problem and possible solutions that take into account multiple aspects (e.g., looks at bigger systems as well as details of solution, recognizes that the solution may involve changes at the societal as well as technological level). This might include examining a combination of factors related to the type of vehicle, the passenger, the design of the seatbelt, manufacturing of the seatbelt, or the frequency of the problem.</td></tr>
</table>

Dimension 2: Teams and Expertise

When putting together a team, what knowledge or skills should the team members have? Are there any specific job titles or positions these people should have? To score this dimension, examine student responses to Question 2.

0	1	2	3
No answer or irrelevant (e.g., smart people).	Mentions teaming skills (e.g., teamwork, communication) or *generic* knowledge, skills, or expertise (e.g., math, scientist).	Mentions *two or more* areas of content expertise specific to solving the problem (e.g., doctor for the elderly) or relevant job titles (e.g., mechanical engineer). Skills related to teaming may be one area of specific expertise	Describes expertise in specific terms and addresses specific teaming skills

Dimension 3: Critical Evaluation of Design

Use student responses to Question 3 to score this dimension. This team did communicate and define the problem (Q3a). They did not conduct any research before developing their foam belt, test the belt, or perform any redesign (Q3b and Q3c). Note that to obtain a score of 2, student can tell you something relevant that the team did well and also a step that the team skipped. They can also get a score of 2 if they don't identify a step that the team did well, but can identify two or more steps that were missed. To obtain a score of 3, students must describe both one step done well *and* multiple steps that were missed.

0	1	2	3
Sees no meaningful need for improvement in design or gives irrelevant responses (e.g., just do more of the same) that are not steps of design process.	Identifies one element of the design process on which the team did well OR sees need for improvement, but focus for improvement is on a single detail or two from the engineering design process.	Recognizes at least one element of the design process on which the team did well AND one specific element or step from the engineering design process that needs improvement or is missing. **OR** Describes need for improvements on multiple steps in the engineering design process.	Can identify at least one element of the design process on which the team did well AND describe need for improvements on multiple steps in the engineering design process.

Dimension 4: Use of Data and Research			
Use responses from Questions 4 and 5 to score this dimension. Students should describe how the data can be used to improve or find a better solution to the problem. Figure 1 compares the force exerted by 3 different types of seatbelts, as well as the force of the airbag and how much the chest is compressed with each seatbelt type. Figure 2 demonstrates that the severity of rib fractures in car accidents increases as adults age. Figure 3 presents the density, level of deformity, and compression of 4 types of materials under force. Examine the responses to ensure that students are not merely restating the titles of the figures. For example, a student should not receive any points for a response of “You can use Table 1 to compare materials under force.” Students may also provide suggestions for research that should be conducted or data that the team might find useful. This should be specific and relevant, not merely “do research on the problem” or “look up answers on the internet”. The students could, however, note that “the researchers should use the internet to see how other teams have tried to solve this problem.” Note that in the second example, the use of the internet is for a specifically stated purpose.			
0	1	2	3
Does not respond or makes an irrelevant statement about the data (e.g, restates the titles of tables/figures without describing how to use the information)	Makes a relevant statement about how to use one element of the data to find a better solution or makes a suggestion about relevant data or information that could be collected.	Makes at least a total combination of two: relevant statements about how to use the data provided to find a better solution, other relevant research to do, or other data to collect.	Provides valid statement(s) about *how to use* all data represented and makes at least one suggestion about other relevant research to do or data to collect.

Appendix C – Informed Consent Form

Title of Project: Understanding Expert Performance on an Engineering Design Task
Principal Investigator: Mary Hibberts, mfk701@gmail.com
Advisor: Dr. R. Burke Johnson, bjohnson@southalabama.edu, (251) 380-2861

Please read this document carefully. If you want a copy of this consent form, you may request one and we will provide it.

This study, conducted by a student in the Professional Studies Department at the University of South Alabama, is concerned with learning more about a series of assessments developed to measure engineering design skills and performance. If you agree to participate, you will be asked to answer a series of engineering design questions related to three engineering design scenarios. Also, you will be asked to provide some demographic information (e.g., your age and gender) and a description of your engineering/design education and past experiences.

Your participation in this research study is completely voluntary. You do not have to participate. You may quit at any time without any penalties. You may also skip any questions on any of the forms that make you feel uncomfortable or discontinue the study at any time. If you agree to participate, the study session will last approximately 90 minutes.

Participating in this research does not guarantee any benefits to you. You will not receive any incentives for participating. However, through participation, you may learn more about how research studies are conducted. This research will help others to learn more about the applied assessments for engineering design performance evaluation. If you wish, you may obtain written information about the outcome of the research by contacting the researcher or her faculty advisor.

To the best of our knowledge, the risk of harm and discomfort from participation is no more than would be experienced daily life. All data from this study will be kept from inappropriate disclosure and will be accessible only to the researcher and her faculty advisor, Dr. R. Burke Johnson. The researcher will use numbers in association with your data to protect your privacy. If your participation in this study has caused you concerns, anxiety, or otherwise distressed you, you can contact the USA Mental Health Center at (251) 473-4423.

For questions about your rights as a research participant in this study or to discuss other study related concerns or complaints with someone who is not part of the research team, you may contact the Institutional Review Board at 251-460-6308 or email irb@southalabama.edu

You have read, or have had read to you, and understand the purpose and procedures of this research. You have had an opportunity to ask questions which have been answered to your satisfaction. You voluntarily agree to participate in this research as described. And, you are at least 19 years of age.

Participant's Signature and Date

USA Institutional Review Board

Approved:	4/01/2015
Expires:	
IRB number:	15-090/733352-1

Appendix D – Demographic and Engineering Experience Questionnaire

Identification Number: ___________

Please answer the following questions about yourself and your experience with engineering/design courses and activities.

1. How old are you?

o 19-25
o 26-35
o 36-45
o 46 or older

2. What is your gender?

o Male
o Female

3. Are you Hispanic/Latino?

o Hispanic or Latino
o Not Hispanic or Latino

4. What is your race? (Check all that apply)

o American Indian or Alaska Native
o Asian
o Black or African American
o Native Hawaiian or Other Pacific Islander
o White
o Other: _________________________________

5. What is your primary language?

o English
o Other:_______________________________

6. What is your major?
___________________________________7.

How interested are you in Science, Technology, Engineering, and Math (STEM) fields of study?

1	2	3	4	5
Not at all interested				Very Interested

8. How interested are you in learning about engineering?

1	2	3	4	5
Not at all interested				Very Interested

9. Did you take any engineering or design related classes in high school?

- No
- Yes (please list the classes you took below)

10. Did you participated in any other activities related to engineering or design in high school (e.g., clubs, training, camps, projects, extracurricular activities, competitions)?

- No
- Yes (please list the activities below)

11. What was the highest level math class you completed in high school?

- Algebra I
- Geometry
- Algebra II
- Trigonometry
- Pre-Calculus
- Other:______________________________

12. What was the highest level science class you completed in high school?

- Biology
- Earth Science
- Physical Science
- Chemistry
- Physics
- Other:_______________________________

13. Did you take any computer or other technology classes in high school?

- No
- Yes (please list the courses you completed below)

14. Have you taken any college courses in engineering?

- No
- Yes (please indicate which areas)
    - Engineering
    - Chemical engineering
    - Civil engineering
    - Mechanical engineering
    - Systems engineering
    - Electrical engineering
    - Computer engineering

Please list the specific engineering courses you have completed below.

15. Have you participated in any other activities related to engineering or design in college (e.g., clubs, training, conferences, projects, extracurricular activities, competitions)?
    - ○ No
    - ○ Yes (please list the activities below)

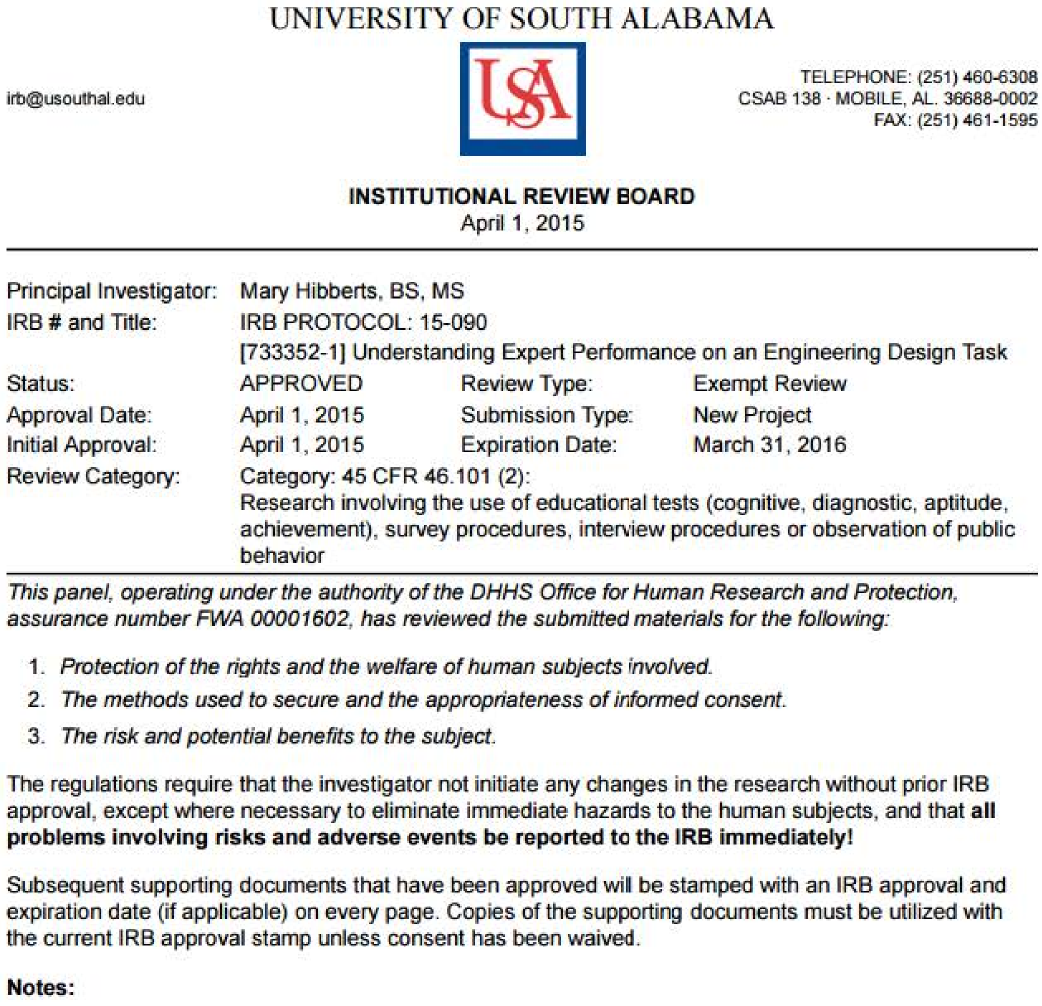
16. Have you taken any science courses at the college level?
    - ○ No
    - ○ Yes (please list them below)

17. Have you taken any design related technology courses at the college level (e.g., computer science)?
    - ○ No
    - ○ Yes (please list them below)

18. Have you taken any mathematics courses at the college level?
    - ○ No
    - ○ Yes (please list them below)

Thank you for completing this questionnaire. Please return this to the researcher along with your completed assessments.

Appendix E – IRB Approval Letter



UNIVERSITY OF SOUTH ALABAMA

irb@usouthal.edu

TELEPHONE: (251) 460-6308
CSAB 138 · MOBILE, AL. 36688-0002
FAX: (251) 461-1595

INSTITUTIONAL REVIEW BOARD
April 1, 2015

Principal Investigator:	Mary Hibberts, BS, MS		
IRB # and Title:	IRB PROTOCOL: 15-090 [733352-1] Understanding Expert Performance on an Engineering Design Task		
Status:	APPROVED	Review Type:	Exempt Review
Approval Date:	April 1, 2015	Submission Type:	New Project
Initial Approval:	April 1, 2015	Expiration Date:	March 31, 2016
Review Category:	Category: 45 CFR 46.101 (2): Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior		

This panel, operating under the authority of the DHHS Office for Human Research and Protection, assurance number FWA 00001602, has reviewed the submitted materials for the following:

1. *Protection of the rights and the welfare of human subjects involved.*
2. *The methods used to secure and the appropriateness of informed consent.*
3. *The risk and potential benefits to the subject.*

The regulations require that the investigator not initiate any changes in the research without prior IRB approval, except where necessary to eliminate immediate hazards to the human subjects, and that **all problems involving risks and adverse events be reported to the IRB immediately!**

Subsequent supporting documents that have been approved will be stamped with an IRB approval and expiration date (if applicable) on every page. Copies of the supporting documents must be utilized with the current IRB approval stamp unless consent has been waived.

Notes:

BIOGRAPHICAL SKETCH

BIOGRAPHICAL SKETCH

Mary F. Hibberts was born in Oak Park, Illinois on February 18, 1984. She graduated from Spring Hill College, with a Bachelor's degree in Psychology and Hispanic Studies in 2006. She also earned a Master of Science degree from the University of South Alabama in Experimental Psychology. She received the Psychology Graduate Student of the Year award in 2009 and the Instructional Design Ph.D. Student of the Year award in 2011. She is an instructional systems specialist at the U.S. Coast Guard Aviation Training Center in Mobile, Alabama. She designs and develops Interactive Courseware for aviation systems training, produces online assessments, and conducts evaluations to validate training and performance outcomes for fixed-wing and rotary-wing pilots and aircrews. Mary has conducted and published research about technology in the classroom, audiovisual speech perception, mixed methods research, and survey sampling techniques. She has a Ph.D. in Instructional Design and Development from the University of South Alabama. Mary is married to Adam Hibberts – they have two children, Helen and John, and a dog named Hank.