ABSTRACT

WARD, ERIK MICHAEL. The Development and Validation of an Instrument to Test the Self-efficacy of Teachers Teaching Engineering Design. (Under the direction of Dr. Aaron Clark and Dr. Cameron Denson).

This study sought to develop and validate an instrument to test the self-efficacy of teaching engineering design. This study is part of a larger effort to develop an instrument for both pre-service and in-service teachers of Science, Technology, and Engineering education. The focus of this study is on the initial instrument creation process and validation with in-service teachers in Technology and Engineering education. A linear regression model found a significant association between the composite score ($F(2,78)=6.469$, $p=.003$, $R^2=.142$) and experience teaching engineering design and general teaching experience as independent variables. Experience teaching engineering design ($t=3.566$, $p=.001$) and general teaching experience ($t=-2.087$, $p=.040$) were both significant predictors, with experience teaching engineering design being significant at a higher level. The dual experience model is preferred over a model with just teaching engineering design as the independent variable ($F(1,79)=8.232$, $p=.005$, $R^2=.083$) although both models show a significant association with the composite score. A one-way ANOVA showed no significant differences based upon gender, although there were significantly different samples sizes between males (n=57) and females (n=24). Levene's test indicated that the homogeneity of variance assumption was met. The initial findings from this study indicate that a case for validity may be made. A second confirmatory study will be needed but falls outside of the scope of this study.

The Development and Validation of an Instrument to Test the Self-efficacy of Teachers Teaching Engineering Design

by
Erik Michael Ward

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Education

Technology Education

Raleigh, North Carolina

2022

APPROVED BY:

Dr. Aaron C. Clark Co-chair of advisory committee	Dr. Cameron Denson Co-chair of advisory committee
Dr. Brian Matthews	Dr. DeLeon Gray

BIOGRAPHY

Erik Michael Ward completed a Bachelor of Science in Technology Education from Ohio State University in 2010. He earned a Master's of Science in Technology Education from North Carolina State University in 2013. After his M.S, he continued in pursuit of his Doctorate of Education in Technology Education. Erik is a member of the honorary society Epsilon Pi Tau. His current academic interests include engineering design, engineering graphics, student learning, and curriculum development.

ACKNOWLEDGEMENTS

I would like to acknowledge thank Dr. Matthew Lammi who started me on this journey. A thank you to each of my committee chairs, Dr. Aaron Clark and Dr. Cameron Denson who have provided guidance and insight throughout the dissertation process and pushed me to do better. A thank you to Dr. DeLeon Gray who provided advice and was always available to answer questions about the statistical analyses. A thank you to Dr. Brian Matthews, who stepped in to fill a vacated committee role; provided assistance and advice to me many times over my journey as a graduate student.

An additional thank you to my family who has always asked about my progress and kept me pushing forward. My mothers, Irene and Kathy, who have provided encouragement and assistance with reviewing and editing.

A very special thank you to my wife, Kady, who has supported me throughout this process and without whom this journey would have been much more difficult. She has listened to hours of my frustrations and trials completing the process, made time to act as a reviewer and editor, and encouraged me to set aside time to research and write even when I struggled.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Introduction

The purpose of this study was to create and establish the validity and reliability of an instrument measuring the self-efficacy of teaching engineering design. This instrument is the self-efficacy of teaching engineering design, or SETED. This instrument takes the concept of self-efficacy and focuses it into the domain of teaching. It then further focuses it from just teaching in general to teaching engineering design. Engineering design is present in science, technology, engineering, and mathematics, STEM, education. This study specifically focused on technology and engineering educators. Engineering design is a key concept in the technology and engineering classrooms, as was evidenced by its presence in the standards for these classrooms (ABET 2012a; ABET 2012b; ITEEA, 2007; ITEEA, 2020). With the push for greater integration across STEM disciplines, interest has risen in teaching engineering design outside of the technology and engineering classrooms, specifically within science classrooms. Engineering design provides a vehicle for integrating technology and engineering concepts across disciplines. It should be noted that technology is not limited simply to computers and electronics but encompasses any way in which humans modify their natural world to satisfy their needs and wants (NAEP, 2014). Science classrooms are already starting to incorporate technology into their classrooms and discussing it in relation to the natural world (Eisenkraft, 2011; Silk and Schunn, 2009).

The Board of Science Educators, BOSE, has included technology, as it related to science and design, into the national standards for science education (2011) and their 2011 *Framework for K-12 Science Education*. Engineering design is also seen in the National Research Council's (NRC) *Next Generation Science Standards (NGSS)*, finalized in 2013. The International

Technology and Engineering Educators Association (ITEEA) also includes engineering design as part of its educational standards for technology and engineering classrooms. This shows engineering design as an important concept, present in the standards of three disciplines within STEM education.

Justification and rationale

Effective education is always a concern for educators and the field of education, and with the inclusion of engineering design in more classrooms, the importance of understanding the effectiveness of the education surrounding it is important. Student achievement finds itself the most widely used measurement for determining the effectiveness of a student's education, and therefore the effectiveness of the teacher. This is evident through the use of statewide testing, and the inclusion of these testing results in teacher evaluations for forty states (Boser, U., 2012). Additionally, a teacher's self-efficacy in teaching a content area shows a correlation to student achievement (Cantrell, 2003). This identifies understanding the teacher's self-efficacy of teaching engineering design a concern when attempting to look at the effectiveness of that teacher in teaching engineering design. While it is important to evaluate the overall effectiveness of an educator, it is important to note that because of how self-efficacy functions it is best to work within a single domain when possible (Zee and Koomen, 2016).

Given the presence of engineering design across multiple STEM classrooms (NRC, 2013; BOSE, 2011; ITEEA, 2007), the researcher was initially interested in understanding the effectiveness of the teachers throughout STEM in teaching engineering design. While student achievement is perhaps the more publicly known measure used for determining this, the relationships between self-efficacy, student achievement, and teacher efficacy identified self-efficacy as a more viable candidate for understanding the effectiveness of educators in teaching

engineering design (Cantrell, 2003; Panadero, Jonsson, and Botella, 2017; Tschannen-Moran, and Hoy, 2007). However, upon searching no instrument existed for evaluating the self-efficacy of teaching engineering design within STEM education. This fact is what gives this study its purpose.

The initial interest in understanding the effectiveness of teachers in teaching engineering design within STEM education implies that science, technology, and engineering classrooms are all of interest. However, engineering design is a relatively new concept to science classrooms. With the inclusion of engineering design in the science standards starting in 2011, finding a large enough population of science teachers with education and experience in teaching engineering design may be problematic. In an effort to assist with making a case for validity, the best course of action was to reduce the population to technology and engineering teachers.

Towards the goal of creating, refining, and validating an instrument to test the self-efficacy of teaching engineering design, the purpose of this study is to refine and validate an instrument designed to test the self-efficacy of teaching engineering design. The goal was to provide a case for a validated instrument, or upon failing to do so, provide a more refined instrument with recommendations that was ready for further refinement and validation.

The limited scope of this study also allows for scalable research with future efforts. The population can be expanded to include pre-service teachers and science education teachers. With or without these population expansions this instrument would allow for future studies to determine not just the effectiveness of teachers, but also the effectiveness of interventions and education related to teaching engineering design through changes in self-efficacy.

Problem statement

Self-efficacy shows relationships to teacher efficacy (Cantrell, 2003; Çakiroglu, Çakiroglu and Boone, 2005; Erdem and Demirel, 2007; Tschannen-Moran and Hoy, 2007), which determines the effectiveness of a teacher in the classroom. An index search of ERIC and Google Scholar returned no instruments which evaluated self-efficacy of teaching engineering design within STEM education. However, it did return several results for performing engineering design and teaching engineering. Given the links between self-efficacy, teacher effectiveness, and student achievement (Cantrell, 2003; Çakiroglu, et. al, 2005; Erdem and Demirel, 2007; Tschannen-Moran and Hoy, 2001), a way to understand the self-efficacy of teachers in teaching engineering design is needed if further research in this area is to continue. This research attempts to fill that gap through the completion of the instrument development process.

Scope of study

Given the available funding and complexity of the process in making a case for validity, this study makes several concessions from the initial interest in the hopes to aid the process. The result is this study focuses on in-service teachers within technology and engineering education. It maintains the focus on the self-efficacy of teaching engineering design. Engineering design is a competency within engineering and technology education given its presence in both ITEEA (2007) and ABET (2012) standards. Several instruments, such as Yoon et al. (2012; 2014), within these fields exist that measure the self-efficacy of teaching engineering. Content and construct validity via subject matter experts, SMEs, along with criterion-related validity will also be used to help strengthen the case for validity (Carberry et al, 2010; Yoon and Evans, 2012; Yoon et al, 2014).

Research Questions

To guide this research the following questions were asked:

Q1. Is there evidence that the SETED scale is reliable?

Q2. Is there evidence of validity of the SETED scale in the theory of self-efficacy?

Q3. Is there evidence that the SETED items represent the domain of teaching engineering design?

Q4. Is there evidence that the SETED scale is a valid instrument?

Limitations

The results of this study will be limited by several factors including:

1. Limited body of prior research in this area requiring the use of low strength analysis.
2. Limited funding restricting options for participant recruitment.
3. Participants will be self-reporting.
4. Participants are volunteers and may represent a limited view of the population.
5. Recruitment for participants took place during the beginning of the COVID-19 pandemic in the United States.

Definitions

Bandura's Social Cognitive Theory **-** a behaviorist learning theory that states learning is affected by performance accomplishments, vicarious experience, verbal persuasion, physiological states, and self-efficacy. (Bandura 1977; Bandura 1982)

Construct Definition **-** The key components of a construct being used to create a framework of understanding. This study uses self-efficacy, which can be broken down into task, ability, and level of belief (Ritter, et al., 2001).

Content Definition - Content definition provides a clear and explicit outline of the content being tested. This definition should break down the content into the individual components the work together to form the overall content. Each component should clearly state how it relates to the content as a whole (Ritter, et al., 2001).

Engineering Design - Engineering design "demands critical thinking, the application of technical knowledge, creativity, and an appreciation of the effects of a design on society and the environment" (ITEEA, 2007).

Scale Selection - Scale selection is the identifying the number of response options for a Likert scale. It also requires providing descriptors above certain points. Scale selection includes not only the terms used as descriptors, but also how many descriptors and which points should have them. (Bandura, 2006)

Self-Efficacy - Self-efficacy is a component of Bandura's social cognitive theory initially put forth in 1977. Bandura (1977) defined self-efficacy as one's belief in one's ability to succeed in specific situations. Self-efficacy determines how successful a person believes they will be at a given task, how long they will continue in the face of adversity, and how much effort they are willing to put forth into the task (Bandura, 2006).

STEM Education - STEM education is an educational approach that integrates Science, Technology, Engineering, and Mathematics classrooms. This approach recognizes that these separate disciplines are interrelated and student learning can be enhanced by relating subject matter across classrooms. (Clark, 2012)

Summary

Engineering design is a standard across multiple STEM disciplines. It is a topic of great importance for understanding how technology is developed and relates to the natural world. The

skills developed through the engineering design process, such as critical thinking and problem solving, are also valuable for every student to learn. Effective teaching helps to increase the transference of this knowledge and the development of skills. One way to help insure effective education is to understand the teacher's teaching self-efficacy for the content. This is due to teacher self-efficacy relating to student achievement. In addition, self-efficacy helps to determine the persistence of teachers in the face of obstacles. By this logic, understanding the self-efficacy of in-service teachers in teaching engineering design becomes important. Towards this end, an instrument designed to test the self-efficacy of teaching engineering design is needed. Unable to find such an instrument to begin research, this study proposes the refinement and validation of said instrument. In order to establish the validity of the instrument, it should be developed for, and tested on, in-service teachers. This is due to the links between self-efficacy and experience (Carberry, Lee, and Ohland, 2010). As well as the availability of similar instruments. This study seeks to develop an instrument for widespread use across the United States focused on in-service technology and engineering education teachers' self-efficacy of teaching engineering design with the possibility for future expansion.

CHAPTER 2: REVIEW OF LITERATURE

Introduction

This chapter gives a review of literature relating to the development of an instrument designed to test the self-efficacy of STEM teachers teaching engineering design. This review begins by looking at the process for developing a self-efficacy scale. Through this literature review several key areas were identified, especially those related to the definition of the construct and content. As a result, this review next covers the basic theory of social cognitive theory, self-efficacy's role in the social cognitive theory, general self-efficacy, and self-efficacy within the domain of teaching. The second part of the construct is that of engineering design. In discussing engineering design, first the reference is set based upon the disciplines in which engineering design is found. This chapter then reviews the specific domain of interest, engineering design. During this review of engineering design, self-efficacy is revisited in a narrower context within the domain of teaching engineering design.

General instrument development methodology

This section examines the process of developing an instrument to measure self-efficacy. Benson and Clark (1982) identify four phases, split into thirteen steps, of creating an instrument. These phases are universal and can be loosely defined as item generation, item refinement, instrument testing, and instrument validation (Baker, et al., 2008; Benson and Clark, 1982; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al, 2001; Smolleck, et al., 2006; Yoon, et al., 2012). Each of these phases can be broken down into more discreet steps, which are explained within the appropriate phase of the instrument development process. Figure 2.1 General instrument development methodology shows the basic methodology of the instrument

development process.

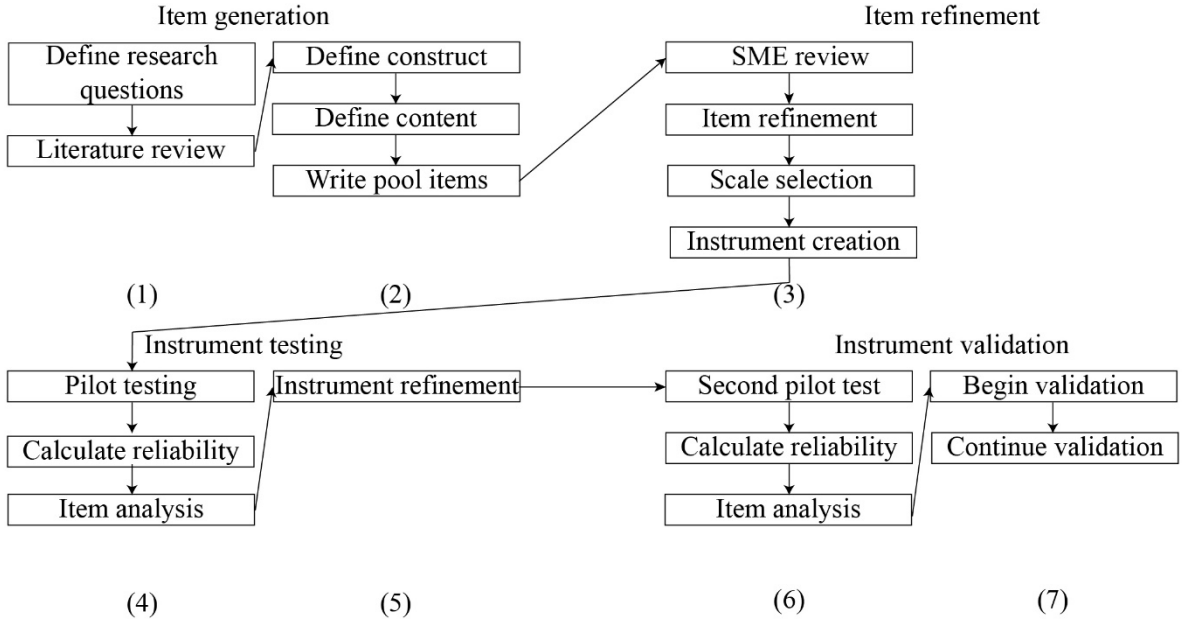

Figure 2.1.

General instrument development methodology

The result of this process is that a large collection of knowledge is filtered down, turned into items, and then filtered further. Figure 2.2 Item count during instrument development, visualizes this process as a funnel, where a large amount of information or items, are placed in at the start and reduced as the instrument becomes more refined. Once an instrument makes it to the pilot testing phase, item reduction is only done if factor analysis or internal reliability testing indicated there is an issue (Dellinger, et al., 2008; Smolleck, et al., 2006). If item reduction occurs, then the process restarts at instrument testing, with the generation of instrument updating to the newest version (Smolleck, et al., 2006).

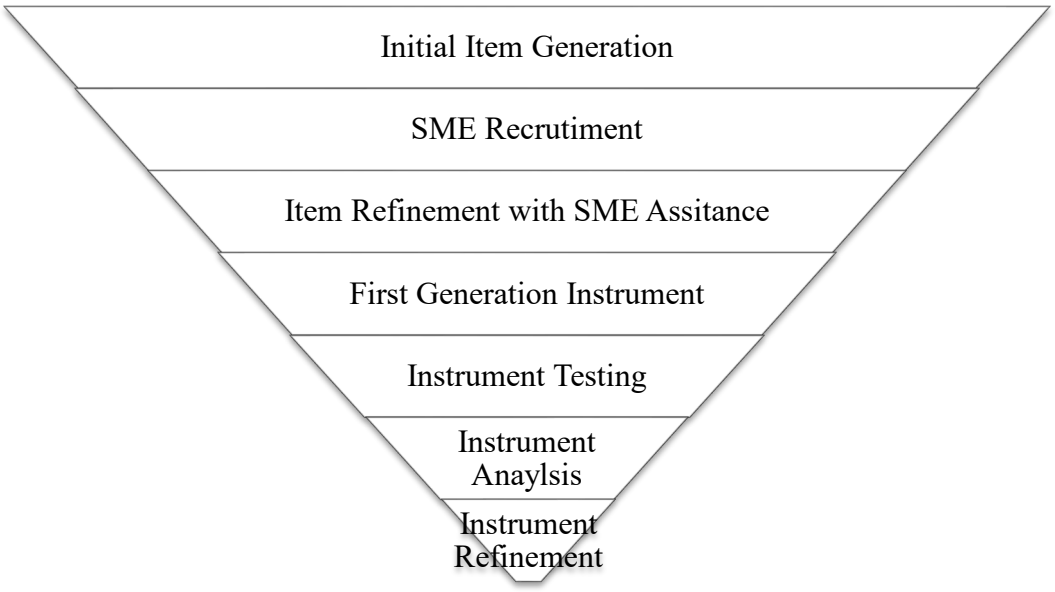

Figure 2.2.

Item count during instrument development

Item generation

Item generation is the first stage of instrument development and breaks into several components: construct definition, content definition, item wording, and scale selection (Bandura, 2006). Before items can be created, a construct is identified as it contributes to the manner in which individual items are worded (Bandura, 2006). Content definition is coupled with construct definition in order to finalize item wording (Bandura, 2006; Dellinger, et al., 2008; Smolleck, et al., 2006). There exist two major methods of generating items. The less common method is seen more in qualitative research and involves asking SMEs open-ended questions, reviewing their responses, and generating items based upon the themes discovered by analyzing these responses. These generated items would then be validated through the SME review process. The more common method involves the researcher creating the items through a literature review and then validating the created items through the SME review process. This second method was used for this study. The beginning of this process is defining the construct.

Construct definition. "When a construct is operationalized, the components necessary to measure it are spelled out." (Benson and Clark, 1982) This can take place in one of two ways, through a review of literature or through a Delphi study. These components then formulate the wording of the individual items or stem. This construct definition allows for validity testing to be performed; however, this operation definition "does not imply that the construct is being measured well" (Benson and Clark, 1982). This issue of quality is addressed later in the validation stages. If the driving construct of the instrument is to be self-efficacy then defining self-efficacy is an early step in the process of item generation (Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al., 2001; Smolleck, et al., 2006; Yoon, et al., 2012). Badura's social cognitive theory (1977, 1995) puts forth the concept of self-efficacy, and thus provides the definition of self-efficacy. Bandura (2006) also states, "Self-efficacy is concerned with perceived capability. The items should be phrased in terms of can do rather than will do" (p. 308). This statement assists with the operationalization of the construct, providing more clear parameters to which wording should adhere. Some researches utilize a stem, which provides a set of base wording expanded upon by the individual items (Bandura, 2006; Dellinger et al., 2008). Other researchers use individual items, with similar or identical wording, that act as complete thoughts to which the subject responds. (Smolleck, et al., 2006; Carberry, et al., 2010; Yoon, et al., 2012) The choice of individual item versus stem usage may not affect the definition of construct; it merely tweaks the eventual wording.

Content definition. Content definition is required before item generation as it combined with construct definition to inform item generation. Content definition involves researching the content in question (Baker, et al., 2008; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al, 2001; Smolleck, et al., 2006; Yoon, et al., 2012). This process identifies the domain of which

self-efficacy is being measured. In the case of Bandura (2006), several examples are given of general teaching self-efficacy. Like construct definition, there are two different processes for defining content, though both involve SMEs.

The methods are separated by where the initial content generation starts. The first method starts content definition with the SMEs. In this method, researchers create open-ended questions about engineering design and administer the questions to experts (Baker et al., 2008). The experts answer the questions and return the results to the researchers. Baker et al, (2008) code this SME feedback; using this coding to generate items. These items are then sent for SME review and checking. This process repeats cyclically until the researchers are left with a finalized list of components returned from all SMEs with very few or no comments.

The second method of content definition keeps the cyclical nature of the first method; however, the researchers themselves generate the initial list of components and submit them to SMEs (Dellinger, et al., 2008; Smolleck, et al. 2006). A review of literature has not shown either method to be clearly superior for the definition of content. The process used is left to the researcher and has not shown negative effects on validity or reliability. Once the process is complete the content definitions are combined with the construct definition to create items (Baker, et al., 2008; Carberry, et al., 2010; Ritter, et al., 2001; Smolleck, et al., 2006; Yoon, et al., 2012).

Item wording. Item wording is the final portion of item generation and is used to separate each facet of the content and test for only one component based around self-efficacy. Literature shows that some researchers use negatively worded items to increase reliability (Baker, et al. 2008); however, the majority of the literature does not utilize negatively worded items (Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al., 2001; Smolleck, et al., 2006;

Yoon, et al., 2012). This implies that the use of negatively worded items does not greatly increase reliability. Additionally, the use of negatively worded items did not adversely affect the reliability of instruments (Baker, et al., 2008), so the use of them is left to the researcher. An additional consideration of item wording relates to the language used. Overly verbose, obscure, or jargon-laden language should be avoided, because this sort of vernacular soon becomes a measure in reading ability and comprehension rather than content (Benson and Clark, 1982). Once the items have been generated through item wording, construct definition, and content definition they are reviewed by SMEs.

Item refinement

Before testing the instrument, each item is sent to a panel of subject matter experts comprised of experts in the content, context, and construct of the instrument. These SMEs provide a review of each item to recheck the construct and content (Benson and Clark, 1982; Carberry, et al., 2010; Ritter, et al., 2001; Smolleck, et al., 2006). If any error is noted on an item, it is adjusted and reviewed again (Benson and Clark, 1982; Carberry, et al., 2010; Ritter, et al., 2001; Smolleck, et al., 2006). Once the items are finalized, they are placed on the instrument with the scale (Bandura, 2006).

Scale selection. The scale for this instrument is a Likert scale (Baker, et al., 2008; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al., 2001; Smolleck, et al., 2006; Yoon, et al., 2012). Research indicates that a process for selecting the number of points on the scale starts with identifying even or odd number of points, selecting the exact number of points, frequency of wording, and then selection of terms (Bandura, 2006; Baker, et al., 2008; Carberry, et al., 2010; Dellinger, et al., 2008; Pajares, 1996; Ritter, et al., 2001; Smolleck, et al., 2006; Yoon, et al., 2012). Research exists for both odd and even number of points; however, majority of the

reviewed literature uses an odd number of points (Bandura, 2006; Dellinger, et al., 2008; Carberry, et al., 2010; Baker, et al., 2008; Smolleck, et al. 2006). This indicates that for this type of instrument the preference should be to utilize an odd number; however, the number of points is left unclear. However, Bandura (2006, p. 312) indicates that even within an odd number scale the central point is moderate assurance. Typically, in odd numbered scales the central point is listed as neutral (Dellinger, et al., 2008; Carberry, et al., 2010; Baker, et. al., 2008; Smolleck, et al., 2006). Bandura (2006) also only adds a descriptor to the most extreme and central values.

Bandura (2006) suggests an eleven-point scale ranging from zero to 100 with 10-unit increments; this can be simplified down to a 0 to 10 scale with increments of one. This 11-interval scale has a greater predictor of performance over a 5-interval scale (Pajares, 1996). This is due to respondents being less likely to utilize extremes, thus shrinking any scale size by two. This scale can be modified during refinement after testing. This refinement is influenced by the distribution of responses. Responses should be distributed over majority of the range.

Scale selection also includes descriptive interval wording. Out of the total number of intervals in a scale, how many intervals should have wording and which intervals should receive this wording is determined. Several researchers identify the extreme intervals (Dellinger, et al., 2008; Carberry, et al. 2010; Baker et al. 2008; Smolleck, et al., 2006; Bandura, 2006). Some researchers identify the center interval as well (Bandura, 2006; Carberry, et al., 2010); the other researchers all use or modified the same instrument, developed by Enoch and Riggs (1990).

Instrument testing

After the items are developed, placed with a marked scale, and refined, pilot testing of the instrument occurs. In pilot testing, the instrument is administered to a sample population (Baker, et al., 2008; Carberry et al., 2010; Dellinger et al., 2008; Smolleck et al., 2006). This sample is

selected as representative of the intended population; however, the selection process is often driven by funding. If a study is meant to be wide spread, the ideal testing subjects are spread throughout the intended population. The instrument is administered to this group and the results are collected (Baker, et al., 2008; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al., 2001; Smolleck, et al., 2006; Yoon, et al., 2012).

At this point reliability and validity testing are performed. There are three major types of validity to consider when looking at validation: content, criterion-related, and construct. Construct validity is required for any test that measures a hypothetical trait, such as self-efficacy. (Benson and Clark, 1982) Criterion-related validity is used for tests that attempt to predict future performance. While self-efficacy is used to help predict future performance, it does not do so through achievement or physical performance testing. This implies that criterion-related validity testing is not required. Content validity is necessary for the instrument, but has been completed before this stage in testing. The involvement of subject matter experts earlier in the process helped to insure content validity.

The results are then tested for construct validity via factor analysis (Benson and Clark, 1982; Baker, et al., 2008; Carberry, et al., 2010; Ritter, et al., 2001; Smolleck, et al., 2006). In a multiple construct instrument, factor loading is performed on the results to test for construct validity (Baker et al., 2008; Carberry, et. al., 2010; Ritter et al., 2001; Smolleck, et al., 2006). If an individual item is found to be crossing constructs, it is removed or reworded. However, this is not the only use of factor analysis; factor analysis is also used to assist with item reduction (Carberry et al., 2010; Smolleck, et al., 2006).

In this second form of factor loading, testing determines if they only tested one factor of a construct. If an item is found to test more than one factor, it is removed or modified (Carberry et

al., 2010; Smolleck, et al., 2006). A third use of factor analysis searches for underlying factors within a construct (Carberry et al., 2010). The second and third uses of factor analysis are often related and assist with item reduction (Carberry, et al., 2010; Smolleck, et al., 2006). In addition to factor analysis internal reliability, testing is performed (Baker, et al., 2008; Carberry, et al., 2010; Ritter, et al., 2001; Smolleck, et al., 2006).

Internal reliability has several options such as the Spearman-Brown coefficient or Cronbach's α. These tests are intended to determine the correlation between all the items and find if they are all measuring the same phenomenon (Benson and Clark, 1982). Cronbach's α is used by Carberry et al. (2010); Yoon and Evans (2012); and Smolleck, et al. (2006) with Spearman-Brown being used by Baker, Krause, and Purzer (2008). Ritter, Boone, and Rubba (2001) uses a coefficient α, but do not specify which one; Dellinger et al. (2008) does not discuss any internal reliability testing during the development process. This presents a strong case for utilizing Cronbach's α for internal reliability.

Instrument refinement. The process of testing and instrument and modifying the items until they properly align with the construct and context is refinement. Refinement is not a finite process (Benson and Clark, 1982; Smolleck, et al., 2006); it is cyclical, using a new sample each time, and continues until the factor loading indicates the items are correctly aligned (Ritter, et al., 2008; Smolleck et al., 2006). The repeated process is sample group testing, factor loading, and then item refinement. The last iteration of this process does not include item refinement. If items are refined, the process needs repeated until they need no adjusting (Benson and Clark, 1982; Carberry et al., 2010; Smolleck et al., 2006). Some studies only complete initial analysis after one round of testing (Carberry, et al., 2010). Any additional actions are addressed as recommendations (Carberry, et al., 2010).

Instrument validation

After instrument refinement is complete, the instrument is placed into a second pilot test. The goal of this pilot test is to validate the instrument. These analyses are a repeat of the previous pilot testing, with one notable exception: the items are no longer refined. The goal of this validation is to demonstrate reliability and validity (Benson and Clark, 1982).

Construct, Context, and Content Validity. The research shows that the first step in validation is defining self-efficacy as a construct and more specifically how the construct found representation in the instrument. Although all of the studies refer to Badura's theories, there are slight variations within these definitions. Dellinger, et al. (2008) utilizes a unique stem they classified as a BELIEF stem. The stem is a base phrasing on which the items expand. In Dellinger, et al. (2008) the belief stem is phrased: "Right now in my present teaching situation, the strength of my personal beliefs in my capabilities to..." (p. 764) which then corresponds to the items. The complete phrase for item 7 of the Dellinger, et al. (2008) instrument reads, "Right now in my present teaching situation, the strength of my personal beliefs in my capabilities to redirect students who are persistently off task is..." (p. 764), and then the respondent selects the appropriate Likert scale value, 1-4, to complete the phrase. Dellinger, et al. (2008) found that "the BELIEF items elicited somewhat different results than the traditional item stems" (p. 756), and that "the BELIEF item stem (responses) were not as strongly correlated with traditional stem responses" (p. 756). They felt this stem is "more consistent with the language of self-efficacy theory" (Dellinger, et al., 2008, p. 756). Construct validation is performed to determine if the items previously developed actually check the construct of self-efficacy (Baker, et al., 2008; Carberry, et al., 2010; Smolleck, et al., 2006). A construct is defined as a behavior that is not directly observable (Bandura, 2006). Given that the instrument is intended to test self-efficacy, a

behavior that takes place internally, it requires construct validity in order for any of the information provided to have merit (Baker, et al., 2008; Carberry, et al., 2010; Smolleck, et al., 2006). It is also performed before the instrument is utilized to ensure that construct validity has not been compromised in the refinement process (Baker, et al., 2008; Carberry, et al., 2010; Smolleck, et al., 2006). Construct validity is then often accompanied by context validity to determine that not only is the instrument sound in the construct of self-efficacy, but it is also rooted in the correct environment (Baker, et al., 2008; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al., 2001; Smolleck, et al., 2006).

Context validation is utilized to determine that the construct has seamlessly been combined with content (Baker, et al., 2008; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al., 2001; Smolleck, et al., 2006). For self-efficacy instruments, validation uses factor analysis to check to see if the items are correlated not only to context specific factors, but to the correct construct as well (Smolleck, et al., 2006). Research relating to specific forms of self-efficacy such as teaching science as inquiry (Smolleck, et al., 2006), engineering design (Carberry, et al., 2010), and the broader concept of teachers' self-efficacy (Dellinger, et al., 2008) all utilize context validation. These studies show in each of their cases that the individual items present on the test relate to the construct of self-efficacy and the subject matter in which self-efficacy has been grounded.

Content validation seeks to show that the subject matter presented in an instrument adequately represents and describes the content in question. Carberry, et al. (2010) utilizes content validation to determine that each of the eight-steps in their engineering design process is represented. They also perform factor analysis to ensure that wording does not over represent any one dimension and that each item exclusively checks the content it was intended to check.

Baker, et al. (2008) provides experts in the field with open-ended questions and then utilizes a thematic count system to generate items that best cover the content being assessed. The result is valid content, directly provided by a variety of experts. Most researchers generate their own content (Bandura, 2006; Carberry, et al., 2010; Dellinger, et al. 2008; Smolleck, et al., 2006), so factor analysis and expert evaluation is later required to ensure that the instrument adequately represented the intended content. This is conducted by researchers submitting the generated items to experts in the field for analysis and comment (Carberry, et al., 2010; Dellinger, et al., 2008; Smolleck, et al., 2006). Research has shown that wording is an influential factor in the results of the factor analysis to check for validity.

Through the review of literature on the development of self-efficacy scales one of the primary areas of concern is the definition of the construct on which the instrument is based. As this study intended to look at developing a scale for the self-efficacy of teaching engineering design, this points to the next areas of required literature review, the definition and contextualization of the construct. This is first done through understand self-efficacy and what the literature offers.

Social cognitive theory

Bandura's social cognitive theory (1977) postulates "the apparent divergence of theory and practice can be reconciled by postulating that cognitive processes mediate change but that cognitive events are induced and altered most readily by experience of mastery arising from effective performance" (p. 191). Bandura (1986) proposed that the social cognitive theory contained three factors that operate under a triadic reciprocal determinism as illustrated in Figure 2.3. Bandura (1986) identified the factors as behavior; personal factors such as cognition and biological events; and environmental influences. In this triadic layout a change in any one of the

factors results in changes in the other two, which in turn each affect change in the other factors (Bandura, 1986). Bandura (1977, 1986) regards human behavior as a cognitive process, with the readiest way to influence this behavior as performance-based. In this form the cognition of the behavior is a personal factor within the triad while the resulting behavior itself is its own factor.

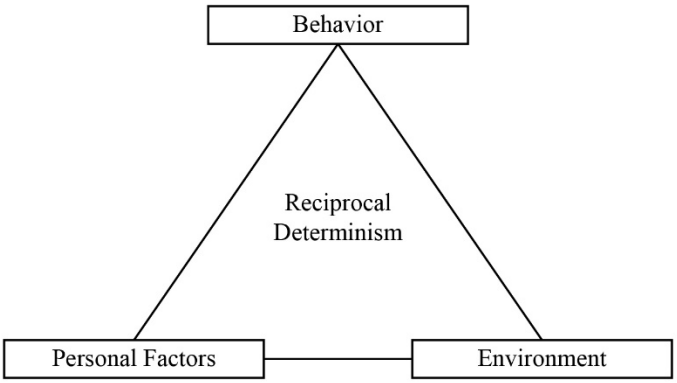

Figure 2.3.

Social cognitive theory triad

"The generalized behavior of an individual is based upon two factors, (a) a belief about action and outcome; and (b) a personal belief about one's ability to cope with a task" (Çakiroglu et al, 2005). Yesilyurt, Ulas, and Akan (2016) define self-efficacy as "a person's own judgement in regarding to realize the capacity to successfully organize the necessary events to achieve the objectives given" and "self-efficacy consists of the regulation of cognitive, social, emotional, and behavioral skills required in order to perform a task and applying effectively to the situation". These two similar definitions show that self-efficacy not only involves the act of regulating behaviors but also the belief in the ability to regulate those behaviors and apply them appropriately. Self-efficacy determines if a person believes they will be successful at a given task; how long they will continue if they encounter an obstacle; how much effort a person puts into the task (Pajares, 1996). These definitions of self-efficacy demonstrate it to be the cognition

which underlies resulting behaviors. As the cognition of behavior self-efficacy represents a key element of the personal factor.

Self-efficacy

Figure 2.4 shows the sources of self-efficacy according to Bandura (1977), "expectations of personal efficacy are based on four major sources of information: performance accomplishments, vicarious experience, verbal persuasion, and physiological states" (p. 195). Falco and Summers (2019) define these four sources as “interpretations of actual performances, vicarious (modeled) experiences, social (verbal) persuasion, and physiological indexes (emotional arousal)”. This definition relays the underpinnings of each of Bandura’s (1977) sources, while also expressing the cognitive nature of self-efficacy.

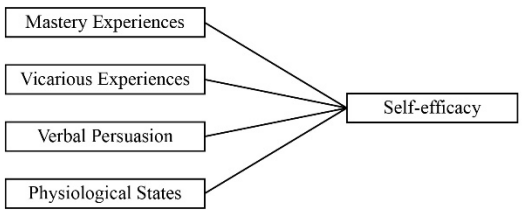

Figure 2.4.

Four sources of self-efficacy

Performance accomplishments revolve around mastery experiences, positive experiences increasing efficacy with negative ones decreasing it. According to Bandura (1977), “After strong efficacy expectations are developed through repeated success, the negative impact of occasional failures is likely to be reduced. Indeed, occasional failures that are later overcome by determined effort can strengthen self-motivated persistence…” (p. 195). This means the overall pattern and timing of failures has an impact on efficacy. Peura, Aro, Raikkonen, Viholainen, Koponen,

Usher, and Aro (2021) states “the manner in which people interpret their past experiences (mastery experience) is the most powerful source of self-efficacy”.

Vicarious experience changes efficacy based on the performance accomplishments of others. These experiences are weaker than those directly gained by personal mastery experiences (Bandura, 1977). Bandura (1977) wrote, “seeing others perform threatening activities without adverse consequences can generate expectations in observers that they too will improve if they intensify and persist in their efforts” (p. 197). These experiences are best gained by observing a model that exhibits a genuine mastery experience of overcoming their difficulties (Bandura, 1977). Additionally, “modeling behavior with clear outcomes conveys more efficacy information than if the effects of the modeled actions remain ambiguous” (Bandura, 1977, p. 197). Peura et al (2021) indicate that the literature has shown weak associations between these vicarious experiences and changes in modern self-efficacy theory.

Peura et al (2021) state that “verbal and social persuasions received from others, such as teachers, parents, and peers, can raise or undermine self-efficacy.” They also define the fourth agent of change in self-efficacy, physiological states, as physiological and emotional states. Peura et al (2021) given an example, “if feelings of stress and anxiety are interpreted as a lack of capability, self-efficacy is undermined”. Outside of mastery experiences being the primary agent of change, it is difficult to identify the importance of the other three sources of changing self-efficacy. Peura et al (2021) report that “although the sources of self-efficacy are theorized to predict self-efficacy development, only a few studies have examined this,” and that these studies “found varying patterns in associations between sources.” Peura et al (2021) go on to postulate that “it is also possible that the relationships between self-efficacy and its sources vary according to the particular skills being developed.” Peura et al (2021) demonstrated the importance of

timing, finding that higher levels of each of the four sources early on resulted in steeper declines of their effects later on. Peura et al (2021) also found an order of the sources to be mastery experiences, physiological states, verbal persuasion, and vicarious experiences.

As self-efficacy is the cognition of behavior, these four sources represent the methods for influencing and changing this cognition within the social cognitive triad. If "people engage in tasks in which they feel competent and confident and avoid those in which they do not" (Pajares, 1996) and self-efficacy is the cognition of behavior, then affecting the self-efficacy of an individual modifies their behaviors. Applying an intervention to any or all of the four sources of self-efficacy that results in a change in any one of the sources causes a chain reaction through the self-efficacy and social cognitive theory that results in behavior and environmental changes for the individual. When this individual is a teacher, these changes are changes in the instructional behaviors and classroom environment.

Zee and Koomen (2016) note that since Bandura's (1997) self-efficacy concept that "scholars started to conceptualize TSE as task or situation specific rather than generalized" additionally citing research was "moving away from the idea that self-efficacy is an omnibus trait, it is acknowledged that TSE beliefs may vary according to different types of tasks, students, and circumstances in class (Ross, Cousins, and Gadall, 1996; Tschannen-Moran et al., 1998)". This shift results in self-efficacy being measured in a more specified rather than generalist way. "Self-efficacy is generally assessed at a more microanalytical level than are other expectancy constructs, which, although they may be domain specific are more global and general self-perceptions" (Pajares, 1996). Concannon and Barrow (2009) note that "in the early 1980s and 1990s, the self-efficacy construct branched into broader measures of self-efficacy such as…", although the examples given represent a large field of self-efficacy research that is focused by

domain rather than a comparatively more macroscopic view. This shows a narrow, domain specific, definition for self-efficacy not only common within self-efficacy research but deemed to be more appropriate. This necessitates that the discussion move from a discussion and understanding of general self-efficacy into a more domain specific context, that of teaching self-efficacy.

Teaching efficacy

Self-efficacy is domain specific (Bandura, 1977; Bandura, 2006) thus when setting the domain of self-efficacy as teaching, one gets self-efficacy of teaching (Zee and Koomen, 2016). However, this domain can be further specified as STEM teaching or along a specific topic, such as engineering design (Tschannen-Moran and Hoy, 2001; Tschannen-Moran and Hoy, 2007). Whereas self-efficacy is one's belief in their actions resulting in a desired outcome (Bandura, 1977), these outcomes are focused upon the person themselves. Teaching efficacy is a belief that judges a teacher's capability to obtain desired outcomes of engagement and learning from students, even those that are unmotivated or difficult (Dussault, 2006; Erdem and Demirel, 2007). In this definition a teaching efficacy reveals a uniqueness, teaching efficacy is not just one's belief in their own abilities to achieve a desired outcome (Bandura, 1977; Bandura, 1982; Bandura, 1986) but to affect change in another person (Dussault, 2006; Erdem and Demirel, 2007) thus changing this secondary party's self-efficacy. When looking at teaching efficacy the focus on these effects is evident.

"In social cognitive theory, efficacy beliefs are considered to be predictors of future behaviors" (Peura, et al, 2021). Teacher efficacy drives a teacher to set higher goals, invest more effort in teaching, and persist in the face of setbacks (Tschannen-Moran and Hoy, 2001) and "previous studies have indicated that teachers' self-efficacy is linked to multiple positive

variables in teaching effectiveness" (Dussault, 2006). Teachers' efficacy beliefs affect classroom management, course structure, and teaching, communicating with students, and influencing student motivation (Erdem and Demirel, 2007). Yesilyurt, et al (2016) defines teacher self-efficacy as, "classroom management, teaching methods, and techniques and use of the computer and instructional tools." Zee and Koomen (2016) also support this definition discussing the movement in definition from Rotter's (1966) theory to Bandura's (1977) theory and then its application by Gibson and Dembo (1984) in development of the Teacher Efficacy Scale which led to Tschannen-Moran and Hoy (2001)'s Teachers' Sense of Efficacy Scale that in turn has led to further domain specific scales. Dussault (2006), notes that general teaching efficacy is defined as "how much teachers believe the environment could be controlled, that is, the extent to which students can be taught, given such factors as family background," whereas personal efficacy "indicates teachers' evaluation of their own abilities to bring about positive change in students". This shows a trend towards specificity within teacher efficacy scales and that teacher efficacy and teacher self-efficacy have become synonymous terms.

Higher instructor efficacy results in students with higher standardized test scores and a positive attitude towards the content (Cantrell, 2003). Yesilyurt et al (2016) found that the teacher self-efficacy was directly related to a student's general academic self-efficacy, content specific self-efficacy, and attitude towards the content. "Research indicates that students generally learn more from teachers with high self-efficacy than those same students would learn from those teachers whose self-efficacy is low" (Cakiroglu, et al, 2005). If a contributing factor to student achievement is effective instruction and education, then determining if effective instruction is taking place is important. Standardized testing is one such method of evaluating the effectiveness of education; higher scoring students have instructors that are more effective.

Instructor efficacy also contributes to student attitude towards content and student achievement (Cantrell, 2003; Yesilyurt, et al, 2016).

Panadero et. Al. (2017) noted that multiple studies have found self-efficacy to be one of, if not the, strongest predictors of academic performance. During a meta-analysis of self-efficacy literature Zee and Koomen (2016) found that many articles only used simple correlations and had low sample sizes with pre-service teachers showing little effect of teaching self-efficacy on instructional practices; however, they found that in-service teachers had a much greater link between teaching self-efficacy and instructional practices. Zee and Koomen (2016) state "although persons may know that certain achievements result in desired outcomes, this information becomes virtually useless when they lack the beliefs that they have the abilities to produce such actions", which is supported by their findings that higher teaching self-efficacy results in teachers being more likely to implement new instructional methods, collaborate with colleagues, and implement data-drive decisions.

Sources of teaching efficacy. Finnegan (2013) identifies the four sources of self-efficacy and contextualize it within teaching. Finnegan (2013) finds that mastery experience is a combination of classroom experience and context, stating "as effective teachers become experts in supporting students to learn, establishing teachers who are motivated in teaching can reduce the problems in education", but "even though teacher self-efficacy is somewhat stable over time, it is influenced by contextual variables". Finnegan (2013) finds vicarious experiences are gathered during pre-service when "teacher self-efficacy appears to be more malleable in novice teachers than veteran teachers", which is consistent with Peura et al's (2021) findings that more experiences earlier in the process yield a greater effect on the shaping of self-efficacy. Additionally Finnegan (2013) cites, "staff development activities introduce teachers to new

teaching strategies" but "teachers rarely observe their fellow teachers at work" and "procedures are rarely in place for teachers to practice the strategies and to receive feedback or coaching". These lack of procedures reduce the ability of in-service teachers to gain meaningful vicarious experiences; although, Dussault (2006) notes that self-efficacy is positively correlated with classroom optimism and willingness to implement new instructional practices implying that teachers, with high self-efficacy, would be willing to do so if given the opportunity. This reciprocal relationship between overall teaching efficacy and the individual sources used to modify it is consistent with the underpinnings of self-efficacy and the social cognitive theories. Conversely, pre-service training provides more of this observation and coaching providing both vicarious and mastery experiences. Some schools have a mentorship program in place allowing for vicarious and mastery experiences for in-service, but these programs end arbitrarily based upon years of experience or similar conditions which relegates their use to novice teachers. Finnegan (2013) finds that the social persuasion and physiological state management sources are intertwined stating "the verbal interaction a teacher experiences about his or her performance and prospects for success from respected others in the teaching context, such as administrators, colleagues, parents, members of the community, etc." and "the level of support an administrator provides to a teacher is one key determinant in their perceptions of teacher self-efficacy."

Thus far the review has discussed general self-efficacy theory and then contextualization of self-efficacy within the teaching domain. However, for this study this domain needs to be further refined into the domain of teaching engineering design. When further contextualizing self-efficacy by narrowing the domain the methods and practice of teaching become known as pedagogy. Before this discussion occurs, contextualizing engineering design through a discussion of the disciplines in which it is found provides a framework for further understanding.

STEM Education

Engineering design is a concept rooted in Science, Technology, Engineering, and Mathematics or STEM education. Science, Technology, Engineering, and Mathematics education have seen changes to their standards that have resulted in more integration (BOSE, 2011; NRC, 2013; ITEEA, 2007). Each area has a set of standards, created by national or international organizations with the input of experts from the field, to drive the educational aim of their respective programs. These standards represent what experts in each area feel students today need to know in order to be successful later in life.

The technology education standards are created by the International Technology and Engineering Educators Association, ITEEA. ITEEA publishes the Standards for Technological Literacy (STL) under the banner of ITEA. ITEEA (2007) defines technological literacy as "the ability to use, manage, assess, and understand technology" (p. 7). ITEEA believes this is important for students so that students "will be comfortable with and objective about technology..." (2007, p. 9). Towards this end, they have created a series of standards that move students towards these goals.

The Board on Science Education, BOSE, created a set of standards for science educators aimed at making students more scientifically literate. The BOSE (1996) defines scientific literacy as "the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity." (p. 22) The BOSE (1996) also states that students should be able to "... express positions that are scientifically and technologically informed." "In learning science, students describe objects and events, ask questions, acquire knowledge…, and communicate their ideas to others" (BOSE, 1996).

Technology, engineering, and design education

Technology, Engineering, and Design (TED) education represents the T and E of STEM education. The American Society for Engineering Education (ASEE, 2013) states part of its purpose is "the advancement of education in all of its functions which pertain to engineering and allied branches of science and technology." ABET's vision states that they "assure quality and stimulating innovation in applied science, computing, engineering, and engineering technology education." This vision identifies the four program areas with which ABET accreditation is concerned two of which are: engineering education and engineering technology education. ABET (2012a) lists general student outcomes for engineering education that include applying mathematics, science, and engineering; identify, formulate, and solve engineering problems; and the ability to design within realistic constraints. ABET (2012b) lists general student outcomes for engineering technology education that include: "the impact of engineering technology solutions in a societal and global context" (p. 3); "ability to select and apply a knowledge of mathematics, science, engineering, and technology to engineering technology problems" (p. 3); and "an ability to apply written, oral, and graphical communication in both technical and non-technical environments" (p. 3.) These two prominent organizations in engineering education provide enough information to generate a simplified definition of the field of technology and engineering education as teaching students to apply math, science, engineering, and technology to solve problems within realistic constraints.

While ABET focuses on post-secondary accreditation, this definition informed the coverage of the discipline. Reviewing ABET (2012a; 2012b) guidelines reveals that engineering and technology go much deeper than this simple definition illustrated. ITEEA (2007) also provides the *STL*, which outlines in detail the topics they felt should be included with technology

and engineering education. There are 20 standards in the *STL* ranging from construction, manufacturing, transportation, to design, and engineering design. ITEEA (2007) also mentions cultural, social, economic, and political effects of technology. While ITEEA's focus was on K-12 education and ABET's on post-secondary, these two groups have very similar ideas for the domain of Technology, Engineering, and Design education. The basic definition generated from ABET (2012a; 2012b) documents holds true and is reinforced by ITEEA's (2007) *STL*. Both of these groups also explicitly mention societal and global effects of technology, indicating that Technology, Engineering, and Design education is a discipline aware of global implications of their field.

In addition to ABET and ITEEA, NRC provides a definition of practices that define Technology, Engineering, and Design education. The *NGSS* stated, "the *Framework* specifies that each performance expectation must combine a relevant practice of science or engineering with a core disciplinary idea and crosscutting concept" (2013, Appendix G p. 1). They further identify eight practices that are essential for students to learn:

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

(NRC, 2013, Appendix G p. 1). These practices are "derived…based on an analysis of what professional scientists and engineers do" (NRC, 2013, Appendix G p. 2). These three organizations help define the domain of TED education. Self-efficacy instruments are used in various parts of TED, being used to assess students (Baker, et al., 2008; Carberry, et al., 2010) and teachers (Yoon, Evans, and Strobel, 2012). Science education notes the relationship between

the areas of STEM through its practices as well as how it views science education and its relationship to the other STEM areas.

Engineering design

Science education recognizes that in order to be scientifically literate students need to understand technology. This is evident in their definition of scientific literacy as well as their standards. The BOSE (1996) has a standard for science and technology, which breaks into three parts: abilities to distinguish between natural objects and objects made by humans, abilities of technological design, and understanding about science and technology. They specifically call out technological design and technology itself. Technology education is responsible for teaching students about technology and making them technologically literate students. In their more recent standards, BOSE (2011) has specifically included engineering design as a separate topic from the aforementioned parts.

The *STL* provides a section for the education of students about technology and design. In this section there is a standard listed as Engineering Design. This concept is important enough to ITEEA to call it out separately so that an emphasis is placed upon learning engineering design. ITEEA identified engineering design for K-12 students by "the engineering design process includes identifying a problem, looking for ideas, developing solutions, and sharing solutions with others" (2007, p100). As a student progresses through their education, this concept becomes more complex than just one simple sentence could convey.

ITEEA is a leading source for the standards and practices of technology and engineering education. Their definition provides an adequate starting point for understanding engineering design; however, engineering design in practice is not clearly definable. Engineering design is however, a process, which means that it does have some order. Engineering design is not trial

and error; it is intentional and purposeful in seeking a solution to the problem (Lammi, 2011). The methods of teaching engineering design are as diverse as its definitions. One of the ways engineering design is taught is through a series of engineering design challenges (Lammi, 2011). These challenges bring in real world examples of how design problems are solved. An understanding of the engineering design problem then informs what practitioners and researchers feel the parts of engineering design were, and which are most vital for students to learn (Lammi, 2011). Lammi (2011) identified five components of the engineering design challenge: relevance, complexness, processes, constraints, and pedagogy. This research was continued and further refined by Lammi, Denson, and Asunda (2018) who used an iterative thematic mapping method to define three themes within engineering design challenges: attributes, habits of mind and practice, and pedagogy. Lammi, Denson, and Asunda (2018) then identified principal attributes for each of these themes. While discussed individually these concepts are intertwined by nature; all being present simultaneously in engineering design problems. However, this nuanced distinction provides a vehicle for further discussion of the engineering design domain.

Challenge attributes

Lammi et. Al (2018) identified the first theme as attributes of engineering design challenges with "four principal attributes of engineering design challenges: relevance, open-endedness, systems perspective, and transdisciplinarity." Lammi et al (2018) state "design challenges can be authentic to engineering and to the student" and "authenticity can help motivate students to explore new learning. In doing so they identify two main types of authenticity, disciplinary and personal.

When defining open-endedness Lammi et al (2018) noted "engineering design is open-ended with respect to the solution as well as the process." This indicates two core points of open-

endedness: the solution and the process. When discussing the process, Lammi et al (2018) stated "Ottino (2004) suggested that the process of design is non-linear, involving multiple decision points that shape and mold over time." This represents a shift from Carberry et al (2010), and Wendell, Wright, and Paugh (2017) who used a linear, step-based approach. Wendell et al (2017) was studying elementary education students where the process was broken into steps for simplicity. But this runs counter to the functioning of engineering design in practice where, "unlike a prescribed procedure, a process tends to be more open and fluid" (Lammi, Wells, Gero, 2020) leveraging the iterative nature of the process to help improve the design as opposed to the linear approach (McFadden and Roehrig, 2019). Chao, Xi, Nourian, Chen, Bailey, Goldstein, Purzer, Adams, and Tutwiler (2017) noted that process is the manner in which an engineering design problem is meant to be solved as well as the creation of artifacts that display evidence of the learner's path. While fluid, this process is a "deliberate activity with a purpose rather than a haphazard or unintentional approach" (Lammi, et al, 2020). In addition to the process, "open-ended activities require the instructor to not have a single right answer." (Lammi et al, 2018) This also means that with multiple solution there may be no best solution (Asunda and Hill, 2007; Jonassen, 2011) or even no known solution. Wendell et al (2017) noted that experienced engineers spend a greater amount of time systematically evaluating tentative designs, reflecting on the problem, and framing the problem. This reflective evaluation is a continual part of the design process (Wendell et al, 2017). As a part of the process, design failures should be expected, how the teacher reacts to these design failures influences the learning of the students (Wendell et al, 2017).

"A system in engineering design has multiple interconnected variables, is loosely bound, involves human factors, and often requires a global or holistic view" (Lammi et al, 2018).

Lammi et al (2018) found that these interconnected variables may be technical, such as mass, or non-technical, such as cost. “These factors can be integrated into K-12 engineering design challenges as constraints” (Lammi et al, 2018). Often constraints such as the laws of physics are included, but these are represented as transdisciplinarity.

The nature of engineering design is that due to engineering, it “transcends other fields of study” (Lammi et al, 2020). Engineering design is readily found in the modern science curriculum and classroom (BoSE, 2011;). “ While engineering design is broad and allows for a myriad approaches and solutions, there are certain mathematical and scientific principles that must be followed” (Lammi et al, 2020). These principles that must be followed are represented as “first principles” (Lammi et al, 2020). Chao et al (2017) found that “teacher(s) monitor students’ progresses and intervene with scientific principles” but that this “strategy is much less effective with larger class size or less experienced teachers.” Additionally, “engineering is not just applied math. Rather mathematical reasoning and skill is a way of thinking and doing.” (Lammi et al, 2018) While Lammi et al (2018) focus on the field of engineering as a whole it is important to note that individually engineering disciplines use engineering design to teach their core concepts (Lammi et al, 2018) and that the inclusion of multiple engineering disciplines can be seen as transdisciplinarity. McFadden and Roehrig (2019) noted that “engineers rarely work in isolation” and it is “important for students to recognize and learn how their own discourse practices can be leveraged to facilitate problem-based discussion when the expected outcome is a team-generated, physical model or prototype.”

Engineering design habits

Lammi et al (2018) found nine themes that fell under their category of “habits of mind and practice” but they also noted “this section does not describe all of the ways of thinking and

doing germane to engineering design." They include a tenth theme towards this idea, cost. Lammi et al (2018) identified their primary principals as: "modeling, graphical visualizations, decision making, problem formulation, questioning, reflection, continuous improvement, optimization and material resources." Wendell et al (2017) noted that "engineering design also requires many other capacities, such as problem definition, skill with technological tools, competence in mathematical modeling, fluency in reasoning about physical quantities, and creative idea generation," all items that can be folded into Lammi's et al (2018) "habits of mind and practice."

Models and graphic visualizations are two of the hallmark artifacts of the engineering design process, finding mention in almost every discussion of engineering design and engineering design problems (Apedoe, et al., 2008; Brophy et al., 2008; Carr and Strobel, 2011; Chao et al, 2017; Cunningham, et al. 2020; Eiskenkraft, 2011; Lammi, et al. 2020; McFadden and Roehrig, 2019; Wendel et al, 2017). Models "can be a tangible prototype, simulation or procedure" (Lammi, et al 2018). In their discussion of teams, McFadden and Roehrig (2019) noted a physical model or prototype, but they also acknowledge the importance of communication and that discourse takes on a variety of forms and practitioners and students must be able to both present and receive information in multiple forms. Engineering design experience differentiates where effort is applied, on the physical model or the underlying concepts and non-physical models (Chao et al, 2017; Wendell et al, 2017).

Engineering design includes decision making, which is a continuous activity throughout the entire process (Lamm et al, 2018; Wendell et al, 2017). Wendell et al, (2017) even noted that the time spent on problem formulation decision making was one of the differentiating factors

between a novice and experienced practitioner, with the novice spending more time doing rather than thinking and analyzing. A part of decision making is the ability to make trade-offs.

"Trade-offs require the designer to make an often difficult decision between opposing variables and solutions." (Lammi et al, 2018) These trade-offs occur mid process, due to engineering designs' cyclical nature, as optimization is sought and may be based upon multiple factors such as material performance and cost (Assunda and Hill, 2007; Brophy et al, 2008; Chao et al, 2017; Carr and Strobel 2011; Eiskenkraft, 2011; Wendell et al, 2017).

"Although there are many different constraints in engineering design, cost is noteworthy." (Lammi et al, 2018). This is especially true in a learning environment where a school must afford the costs of engineering design problems. Access to funding, and as a result materials, changes how engineering design problems look and are experienced by students (Chao et al, 2017).

Pedagogy

Lammi et al (2018) identified four principals of their pedagogy theme: initial student reaction, scaffolding, instruction preparation, and assessment. Pedagogy is not a function of engineering design itself, but it is an important part of the engineering design problem. Since engineering design problems have been used to understand engineering design, it was important to include in this discussion.

"Too often students are accustomed to assignments or problems with one solution. The engineering design challenges fly in the face of most students' previous experiences" (Lammi et al, 2018). This becomes problematic for students when trying to solve engineering design problems as they lose their frame of reference based upon past experiences. With the inclusion of engineering design at earlier levels and across disciplines as indicated by the BoSE's standards

and studies such as Chao et al. (2017) and Wendell et al. (2017), in time this shock may subside as it becomes part of the mainstream education and no longer "antithetical to traditional learning experiences for students." (Lammi et al, 2018). To ease this transition scaffolding, or temporary instructional aide, is helpful for students; although, this scaffolding may be based upon factual knowledge or the engineering design process itself (Lammi et al, 2018; Wendell et al. 2017). Since the instructor will have to provide this scaffolding, their ability to do so is important. Instructors not prepared to provide the appropriate scaffolding as students encounter difficulty with the process, factual knowledge (such as first principles), or dealing with failure can result in negative outcomes for the student (Aydeniz, Bilican, Senler, 2020; Chao et al, 2017; Lammi et al, 2018; Wendell et al, 2017). "Failure to discuss their shortcomings with other practitioners and mentors...may hinder development in teacher's instructional practices," (Lammi et al, 2018) and this failure may be a result of their self-efficacy as discussed earlier in this chapter. Lammi et al, (2018) note that many of these pedological principles need further research. One aspect of the pedagogy that is less ambiguous is assessment. This includes the assessment of artifacts, a concept discussed previously.

Assessment. Evaluation is a vital component of teaching engineering design and the engineering design problem. Chao et al. (2017) noted that the teacher's ability to intervene differentiated novice and experienced teachers. This intervention would take place based upon formative evaluation, whether formal or informal. Chao et al (2017) also found in their literature review that "students' oversight on scientific principles may be attributed to the lack of feedback". Feedback during the instructional process is done through an evaluation process. The most prominent form of evaluation, and thus feedback, are rubrics (Lammi et al, 2018).

Metrics create an expectation for students of a performance standard. Elliot, Murayama, and Pekrun (2011) state,

> " task-based goal pursuit may be more prevalent in classrooms utilizing an absolute grade distribution…whereas self-based goal pursuit may be more common in …settings where intrapersonal improvement is emphasized or even incorporated into the evaluation process."

The implication is that the structure of the course, and the metrics used, can change and direct the focus of the students. This raises the importance of the metrics used to evaluate engineering design problems.

Education uses two main types of evaluation, formative and summative. There is also formal and informal. When discussing metrics, rubrics, and evaluations the concept of an assigned grade is often what comes to mind. This is formal assessment and often summative. The feedback discussed by Chao et al (2017), is informal and should "be specific and relevant" (Lammi et al, 2018). If it occurs during the process, rather than at the end it is formative. Exams and capstone projects are examples of summative evaluation. The intentional, purposeful use of each of these forms of evaluation along with the creation of metrics provides the structural framework of the educational experience of the engineering design problem.

Having looked at instrument development methodologies and the construct of instrument this discussion concludes by looking at the standards laid forth for any psychological testing and evaluation. These standards are generalized for any instrument, not just self-efficacy scales, and present a benchmark for which any instrument should strive.

AERA standards

The American Educational Research Association, AERA; American Psychological Association, APA; and National Council on Measurement in Education provide *Standards for educational and psychological testing* with the purpose to "provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (AERA, 2014); however it also states "the *Standards* should not be used as a checklist" and "evaluating the acceptability of a test or test…does not rest on the literal satisfaction of every standard in this document". The implication is that standards appropriate to the test being developed should be addressed and evidence provided addressing each of these standards. The AERA (2014) is broken into several chapters, within these chapters exist 13 sections, each of which provides a set of standards for tests. Only the first four sections are relevant to this study, as further sections involve the use of an established instrument. Each of these sections has twenty or more standards associated with it. As the AERA (2014) guidance within the *Standards for Educational and Psychological Testing* indicates, the standards are not to be treated as a checklist so listing each of the standards and discussing them would also be inappropriate. However, the selection of a subset of standards to which the instrument being developed should adhere and discussing these standards would be appropriate. Further discussion within this section will identify each of the standards that should be most relevant to this instrument. A discussion of each of the standards presented in the proceeding sections will be presented along with the discussion of the evidence provided towards that standard within the discussion of Chapter 5.

Validity standards

The *Standards for Educational and Psychological Testing,* hereafter referred to as *Standards* begin with a discussion of validity. There are four sources of validity, which intentionally "does not follow historical nomenclature" (AERA, 2014) these sources are test content, response processes, internal structure, and relations to other variables. The first set of standards relate to the clear articulation of each intended test score interpretation and the validity evidence to support each interpretation. Standard 1.0 is the clear articulation of each intended test score interpretation along with the related evidence. Standard 1.1 states the intended interpretation and use, intended population, and construct or constructs should all be clearly stated. The relevant guidance given for this standard is "specify in clear language the population for which the test is intended, (and) the construct it is intended to measure" (AERA, 2014). Standard 1.8 is the reporting of as much detailed information as possible on the sample population from which validity evidence was obtained. Some of the guidance is related to the use of tests outside the scope of this study's scope. The relevant guidance provided is "when the sample is intended to represent a population, that population should be described and attention should be drawn to any systematic factors that may limit the representativeness of the sample" (AERA, 2014) Standard 1.9 relates to the presentation of the qualifications of the experts used for validation, any training or instructions provided, the independence of decisions, and the level of agreement reached. Several pieces from the guidance on this standard are important within the scope of the study namely "systematic collection of judgments or opinions may occur at many points in test construction" and that "whenever such procedures are employed, the quality of the resulting judgements is important to the validation". (AERA, 2014). The guidance also notes that, "the basis for specific certain types of individuals… as appropriate experts…should be

articulated" (AERA, 2014). Standard 1.10 is the reporting of conditions in which data collection occurred when this data is used for statistical analysis, especially those conditions that differ from the intended collection process. The relevant guidance for this study relates to the mode of testing administration and time allowed for test takers to respond (AERA, 2014). Standard 1.11 is the rational for test score interpretation, the appropriateness of content, and procedures to generate content. The guidance suggests "a logical structure that maps the items on the test to the content domain" (AERA, 2014). Standard 1.13 is the rationale for interpretation based upon the relationship among test items and the internal structure of the instrument. "It may be claimed, for example, that a test is essentially unidimensional. Such a claim could be support by multivariate statistical analysis, such as factor analysis" (AERA, 2014). Standard 1.14 relates to the development of subscores and composite scores and the interpretation of differences. The guidance provides helpful direction here with, "the basis for combining scores and for how scores are combined (e.g., differential weighting versus simple summation) should be specified" (AERA, 2014). Standard 1.15 discusses interpretation about the performance on individual items or subsets of items. The helpful information provide here is that "test manuals and score reports should discourage overinterpretation of information." Following validity, the standards move to those for reliability and precision of measurement.

Reliability standards

Standard 2.3 requires the reporting of reliability for each score or subscore upon which interpretation is presented. The relevance section of the guidance is "users should be supplied with reliability data for all scores to be interpreted" (AERA, 2014). Standard 2.5 requires that reliability estimation procedures should be consistent with the structure of the test. Specifically, "a single total score can be computed on tests that are multidimensional" and these scores

"should be treated as a composite score" (AERA, 2014). Standard 2.19 discusses the method of quantifying the reliability scores, the procedures used to select test takers from the sample, and a description of those selected. Namely, "information on the method of data collection, sample sizes, means, standard deviations, and demographic characteristics" (AERA, 2014).

Fairness standards

Standard 3.1 is the design for all steps in the process promote use in the widest range of individuals as possible. "Test developers must clearly delineate both the constructs that are to be measured by the test and the characteristics of individuals and subgroups in the intend population of test takers" (AERA, 2014). The *Standards* guidance further discusses principles of universal design things such as avoiding test speededness; simple, clear, and intuitive testing procedures and instructions; and precisely defining constructs. Standard 3.2 is the minimizing the effect of construct-irrelevant characteristics. According to the guidance this means "use language in tests that is consistent with the purposes of the tests" and "the level of language proficiency…should be kept to the minimum required…to represent the target construct" (AERA, 2014). Standard 3.3 is the inclusion of relevant subgroups in preliminary studies when constructing the instrument. This means including "individuals from relevant subgroups…in pilot or field test sampes used to evaluate item and test appropriateness for construct interpretations" and "expert and sensitivity reviews can serve to guard against construct-irrelevant language" (AERA, 2014). Standard 3.4 is comparable treatment during test administration and scoring. The main concern to be addressed is "computerized and other forms of technology-based testing add extra concerns" such as "working on older, slower equipment" and the "speed of processing or movement from one screen to the next" (AERA, 2014). Standard 3.13 is "a test should be administered in the language that is most relevant and appropriate to the test purpose" (AERA, 2014). Standard 3.17

is evidence for reporting comparable scores across subgroups. The guidance here states "reporting scores for relevant subgroups is justified only if the scores have comparable meaning across these groups"; additionally, "terms used to describe subgroups are clearly defined, consistent with common usage" (AERA, 2014).

Instrument development standards

Standard 4.1 is the definition of construct, definition of the intended population, and interpretation of results. Namely "usefulness of test interpretations depend on the rigor with which the purpose(s) of the test and domain represented by the test have been defined and explicated" and that the domain "should be sufficiently detailed" (AERA, 2014). It should be noted that this standard is specifically referring to test specifications. Standard 4.7 is the procedures for the development and selection of items in the item pool. "The qualifications of individuals developing and reviewing items and the processes used to train and guide them in these activities are important aspects of test development documentation" (AERA, 2014). Standard 4.8 is the use of empirical analyses and the use of expert judges along with the qualifications of the selected judges. "Expert judges may be asked to…identify material likely to be inappropriate, confusing, or offensive" and that these experts are "independent of the test developers" and "judge the degree to which item content matches content categories…and provide balanced coverage of the targeted content" (AERA, 2014). Standard 4.9 is selection of a sample that is representative of the intended population. This means "item and test form tryouts should include relevant examinee groups," and differences between groups should be analyzed (AERA, 2014). Standard 4.12 is the domain of the final instrument represents the defined domain. "Developers should provide evidence of the extent to which test items…represent the defined domain" and "such evidence may be provided by expert judges." (AERA, 2014).

Standard 4.17 is the statement about use for research only or operational use. This standard is included since "this standard refers to tests that are intended for research use only" but "does not refer to standard test development functions" (AERA, 2014). As this study is part of a dissertation, the resulting instrument sits somewhere between these two areas with leanings towards the former. Standard 4.22 relates to the procedures for interpreting scores. Namely, whether scores "should be interpreted as indicating an absolute level of the construct being measure or as indicating standing on the construct relative to other examinees, or both" (AERA, 2014). Standard 4.23 relates to a score derived from the use of weighted items and the rationale of the weighting system used. AERA (2014) provides this relevant guidance, "in many cases, content areas are weighted" and "the rationale for weighting the different content areas should be documented".

Summary

This chapter looked at the construct of self-efficacy, the domain of engineering design, and the discipline of STEM education. Self-efficacy is a part of Bandura's (1977) social cognitive theory. The self-efficacy of the teacher is one of the largest predictors of student achievement (Cantrell, 2003) and even influences a teacher's professional behaviors (Tschannen-Moran and Hoy, 2001; Zee and Koomen, 2016). With student achievement becoming part of the evaluation of professional teachers (Boser, 2012), increasing the self-efficacy of teachers helps to create more effective teachers (Cantrell, 2013). Self-efficacy is domain specific (Bandura, 2006; Zee and Koomen, 2016), so in order to understand the self-efficacy of a teacher a domain must be defined. While there does exist a general domain of teaching efficacy, engineering design is present across the standards for many STEM disciplines (ABET 2012a; ABET 2012b; ITEEA, 2007; ITEEA, 2020; NRC, 2013) and represents a more specific domain than that represented by

teaching efficacy. Understanding these theories plays an important role in the development of an instrument to test the self-efficacy of teaching engineering design. This chapter began with looking at the literature around the development of self-efficacy scales. It identified the general methodological process used in test development, discussed key inflection points of decision making, and ended with a presentation of standards for test development and administration.

CHAPTER 3: METHODOLOGY

Introduction

Benson and Clark (1992) identified four phases of instrument development split into thirteen smaller steps. While the individual steps vary, the same four phases: item generation, item refinement, instrument testing, and instrument validation, are part of the instrument development process. (Baker, et al., 2008; Benson and Clark, 1982; Carberry, et al., 2010; Dellinger, et al., 2008; Ritter, et al, 2001; Smolleck, et al., 2006; Yoon, et al., 2012). Within this general methodology there are several decision points which influence the exact methodology used to develop the instrument. This chapter begins by reviewing two guiding studies and the methods they used for generating their self-efficacy instruments. This chapter then outlines the methodological process used for this study, which used the methods from both of the guiding studies as a foundation. Additional studies provided further support to this foundation when the methodological framework was established. The goal of this study was to develop and pilot test for validation a scale measuring the self-efficacy of teaching engineering design. This goal resulted in several research questions being created, which were:

Q1. How well do the items in the instrument represent the teaching engineering design task in eliciting the self-concept of self-efficacy?

Q2. How well does the instrument predict differences in self-efficacy held by individuals with a range of experience teaching engineering design?

Q3. Is there evidence that the SETED items represent the domain of teaching engineering design?

Q4. Is there evidence of validity of the SETED scale in the theory of self-efficacy?

Previous Studies

There are two example studies that illustrate similar, but different, methodologies in the instrument development process. These studies have been selected because of their relationship to engineering, teaching, engineering design, and self-efficacy. The first is a 2012 study by Yoon, Evans, and Strobel in which they developed an instrument for teaching engineering self-efficacy. Their study focused on content and face validity processes.

The second study was a 2010 study by Carberry, Lee, and Ohland in which they developed an instrument focused on the self-efficacy of performing engineering design. Their study focused on content, face, and concurrent validity processes. This study will closely follow the methodology presented by Carberry et al (2010), utilizing the Yoon et al (2012) study to provide additional support for certain methodological practices.

Yoon and Evans study

In the creation of their instrument to measure teaching engineering self-efficacy, Yoon and Evans (2012) outline eight steps in their research process. This study was followed up by Yoon, Evans, and Strobel (2014), which continued the process of instrument development and validation. As a result, both these concurrent studies were viewed as a continuous effort and study. The first step is a review of literature about the creation of teacher self-efficacy instruments. The second step was a literature review of professional development for the intended population. The third step was the generation of items based upon the literature review. This also included the refinement of items to reduce redundancies, positive wording of items, and exchanging word choices to match items. The fourth step was a review of the items by a panel of SMEs. The fifth was determining the format of the instrument based upon the literature

review. The sixth step was participant recruitment and data collection. The seventh step was data analysis. The eight, and final, step was result reporting.

Step one through three: item development. The first three steps focus on the creation of items. Yoon, et. al. (2012) use several instruments in order to construct their instrument, allowing for multiple constructs within the instrument. "For consistency and clarification, item redundancies were eliminated and all items were rephrased to be statements, not questions, to be positively worded (e.g., "I can" instead of "I can't", and to eliminate inconsistencies in word choice (e.g., "student", instead of "child")" (Yoon and Evans, 2012). These steps are done in order to create the initial items of the instrument.

Step four: item review. Step four is the inclusions of the subject matter experts and their review of the intended instrument. Yoon and Evans (2012) indicate the use of "a panel of professors and graduate students in engineering and education disciplines". Yoon and Evans (2012) indicate that each item is evaluated based upon level of confidence that the item aligned with a set construct. Their study used multiple constructs, and this process was used to help ensure that each item only aligned to a single construct.

Step five: instrument creation. The fifth step could be considered compiling the instrument. In Yoon et. al. (2012) this focused on the format "determined using the suggestions about improvements of teacher self-efficacy instruments for future study in the literature". This included scale selection and labeling of points. This involves taking the items generated in the previous four steps and placing them in the instrument. Yoon et. Al. (2012) then use a Likert-type scale for respondents to indicate how much they agree or disagree with each particular item. Yoon et. Al. (2012) use a six-point Likert type scale. They do this based upon previous research

which indicated that there should be no neutral center point. In addition, they elect to label each of the six points from strongly disagree to strongly agree.

Step six: data collection. This step is the recruitment of participants and data collection. Yoon et. al. (2012) utilized a web-based survey software, Qualtrics, to construct and host the instrument online. They next recruited teachers in their intended population via email. Their study had 153 participants and gathered demographic data. They felt this was an adequate number in order to perform their various analysis tasks.

Step seven: data analysis. The seventh step of this study is data analysis. The goal of this step is to provide answers to each of the research questions. Yoon et. Al. (2012) employ exploratory factor analysis and Cronbach's α for reliability. The data gathered in the Yoon et. Al (2012) study contained data that was not normally distributed, nor was it considered ordered. As a result, the exploratory factor analysis, EFA, performed was based upon weighted least squares. Eigenvalues greater than one were kept, resulting in seven possible factors. Yoon et al (2012) "based on Stevens' (2002) guideline about the relationship between the sample size and cutoff factor loading" generated .40 as the cutoff value. Any items loading onto more than one factor was excluded, and any item not significantly loading onto a factor was discarded. Through the loading process only six factors were used, and the items reduced to forty-one. Yoon et al (2012) then performed reliability testing on each factor.

Cronbach's α was calculated for all items as well as each factor found through the EFA process (Yoon et. al., 2012). Cronbach's α values were between .837 and .977 so it was determined that the instrument was internally consistent. Yoon et al (2012) also note that the removal of any one item did not increase Cronbach's α, implying that each item contributed to the overall strength of the instrument.

Carberry, Lee, Ohland study: Measuring Engineering Design Self Efficacy.

The goal of Carberry, et. Al (2010) was to design and validate an instrument to test the self-efficacy of performing engineering design. They sought to answer three research questions:

"(a) how well the items in the instrument represent the engineering design process in eliciting the task-specific self-concepts of self-efficacy, motivation, outcome expectancy, and anxiety, (b) how well the instrument predicts differences in the self-efficacy held by individuals with a range of engineering experiences, and (c) how well the responses to the instrument align with the relationships conceptualized in self-efficacy theory."

In order to do this, they developed a 36-item online instrument. Carberry, et. Al (2010) do not identify a clear methodological process; however, one can be deduced from the information provided. Carberry et. Al. (2010) seemed to use the process outlined in the following section.

Steps one and two: item creation. The first step was defining the content. Carberry et. al. (2010) knew they wanted to study engineering design. However, they note that "there is no consensus on what exactly constitutes the engineering design process." Therefore, they resolve to use a pre-established process provided by the Massachusetts Department of Education, MDOE. The second step was generating items from this pre-established process. An item is created for each step of the process and for each of the four constructs in their study. This results in thirty-two items. The remaining four items come from an item explicitly questioning conducting engineering design. Carberry et. Al (2010) note the importance of establishing content validity in these steps. They discuss the methods used by previous studies. They identify various practices that are presented in earlier chapters of this document. Carberry, et. Al (2010) seem to conclude that no matter the method used, as long as the content has been reviewed and approved by

subject matter experts it can be considered valid. This is the reason behind the use of the MDOE model, it was created through the use and review by subject matter experts.

Step three: instrument creation. The next step was to take the 36 items and place them alongside a scale. Carberry, et. Al (2010) use a 0-100 scale with points occurring every 10 units. They also identify each of the ends as well as the centermost scale point. The items were grouped into groups of nine, each group focusing on one of the four self-concepts from the research questions. This provided the foundation of the instrument.

Step four: data collection. The fourth step was data collection. An unspecified online survey tool was used, and participants "were solicited … through email listings available to the researchers," (Carberry, et. Al, 2010). These efforts gained three hundred and sixty-seven respondents who were filtered down to two hundred and two for various reasons (Carberry, et al., 2010). These reasons included answers that did not vary or respondents that did not completely finish the instrument (Carberry, et. Al., 2010).

Step five: data analysis. The fifth step was data analysis. Carberry et al (2010) used several different analyses in order to answer each of their research questions. They conducted reliability, factor analysis, correlation, and analysis of variance tests. Cronbach's α was used to determine inter-item reliability for each of the four self-concepts; additionally, these tests were performed by gender to "ensure that the reliability of the instrument was not affected by a respondent's gender." (Carberry, et. Al, 2010)

Exploratory factor analysis was used for each of the four subsets, and revealed one factor per subset (Carberry, et. Al. 2010). In addition, confirmatory factor analysis was done across each subset for each individual item to ensure item consistency (Carberry, et al. 2010). For each of these processes only factors with eigenvalues greater than one were used. These tests only

used the eight items generated from converting the MDOE engineering design process into items.

The final four items, directly asking about engineering design, were used in correlation tests. A calibrated score was generated through the average of the other eight responses. This score was then compared to the engineering design item, to determine if they were correlated. Pearson correlation was used, and Carberry, et. Al (2010) found them to have significant positive correlations which suggested "responses were consistent between ED scores and EDP factor scores".

One-way analysis of variance, ANOVA, compared the composite average scores of the eight items to respondent's experience in engineering (Carberry et. Al, 2010). To do this, Carberry et. Al. (2010) broke respondents into three general categories of experience, those with little to no experience, current learners with some experience, and those with first-hand experience. Each of these categories was then labeled as low, middle, or high self-efficacy in turn. Carberry, et. Al (2010) rely on the assumption that those with more experience will have higher self-efficacy. They use this assumption along with their distributed population in the one-way ANOVA. The ANOVA is done between the average score of each subset and the respondent's experience level. Tukey HSD was done post hoc to determine if the results were significantly different for each group (Carberry, et. al, 2010). The same analysis was done relative to the single item response related to performing engineering design.

Carberry et al (2010) also created hypothesized responses for each of the experience groups and compared these responses to the average response from each group. This was done as a source of construct validity, to show that the sample responses aligned with the expected theoretical responses. In addition, Carberry et al (2010) also correlated the responses to the self-

efficacy items with the items of the other three constructs. These showed significant correlations that aligned with the theoretical relationship between the constructs.

This study

This study used the methods placed forth by Carberry, et al. (2010), with support from self-efficacy instrument development literature as needed to provide extra data points when making a case for validity of the SETED instrument. The previous section discussed the five steps used by Carberry et al., (2010) in their study. This study uses four phases of instrument development, each with underlying steps that are consistent with the literature review. These phases are identified as: item generation, item refinement, instrument testing, and instrument validation. Figure 3.2 illustrates the methodology used in this study.

Item generation

Item generation was a two step process, each of which had several sub-steps. The first step was research scope which identified the research questions and performed a literature review. The second step was the creation of an initial item pool which included defining the construct, defining the content, and writing the pool items.

Research scope. This research wanted to look at the self-efficacy of teachers teaching engineering design. The first question generated was: what is the self-efficacy of STEM educators in teaching engineering design? The second question was: are there any significant differences in self-efficacy of teaching engineering design based upon experience? The final question was: are there any significant differences in teaching engineering design based upon discipline? A review of literature found that no instrument existed to test the self-efficacy of teaching engineering design. As such, developing such an instrument was the required first step. The result of this initial literature review was a change in scope of this study from answering the

original research questions to developing an valid instrument to answer these questions. The development of a valid instrument for determining the self-efficacy of teaching engineering design would then have to be valid within the constructs of self-efficacy and teaching engineering design. This left four research questions for this study to answer:

Q1. Is there evidence that the SETED scale is reliable?

Q2. Is there evidence of validity of the SETED scale in the theory of self-efficacy?

Q3. Is there evidence that the SETED items represent the domain of teaching engineering design?

Q4. Is there evidence that the SETED scale is a valid instrument?

After generating these research questions, a second review of literature was performed. This literature review was the beginning of the second step, generating an initial item pool.

Writing pool items. In order to create an item pool, the literature review focused on defining the construct of self-efficacy and the concept of teaching engineering design. This literature review can be found as a part of the previous chapter. Based upon this literature review, a pool of items was generated (Kukul and Karatas, 2019; Malandrakis, Papadopoulou, Gavrilakis, Mogias, 2019; Sun and Rogers, 2020; Tsai, Wang, Wu, Hsiao, 2021; Xu, Williams, Gu, 2019; Yoon, Evans, and Strobel, 2012; Yoon, Evans, and Strobel, 2014). This first step of generating items was formation of a stem that was consistent with self-efficacy theory. Using Bandura's (2006) guidance for stem creation, two stems were created: "I believe I can teach students" and "I feel I can teach students". The second step to creating items was to generate the content which would be paired with each of the stems. During the review of self-efficacy it was found that self-efficacy is domain specific, so the domain of teaching engineering design was reviewed. This literature review was also presented in the previous chapter. Using this literature

as guidance, sixteen important characteristics of teaching engineering design were identified. Items were then generated based upon these characteristics. The items intentionally included duplicates for some of the characteristics, while other characteristics required multiple items to fully express the content contained within. The result was a pool of forty-one initial items.

Item refinement

Item refinement contained only a single step that was further broken down into four sub-steps: SME review of pool items, item refinement based upon this review, scale selection, and instrument creation. The SME review and item refinement process were cyclical until no further modifications were required (Ritter, et al, 2001; Smolleck, et al, 2006).

SME review and item refinement. The initial item pool was sent to subject matter experts in the fields of teaching engineering design and self-efficacy (Kukul and Karatas, 2019; Malandrakis, et. Al., 2019; Sun and Rogers, 2020; Tsai, et. al., 2021; Xu, et. al, 2019; Yoon, et al., 2012; Yoon, et al, 2014). These SMEs were all faculty members at different universities. The items were placed on a form for review and sent to each SME. The SMEs were asked to respond to each item individually. There was also a question for general or further feedback that pertained to the entire instrument. Through the SME process several items were changed or added. The advised changes were reviewed and acted upon by the researcher. The new items were then sent back to the SMEs for another round of review. This process continued until the items no longer needed modification. The result of this process was items were reduced from forty-one to twenty-eight. These twenty-eight items were then combined with a scale in order to create the SETED instrument.

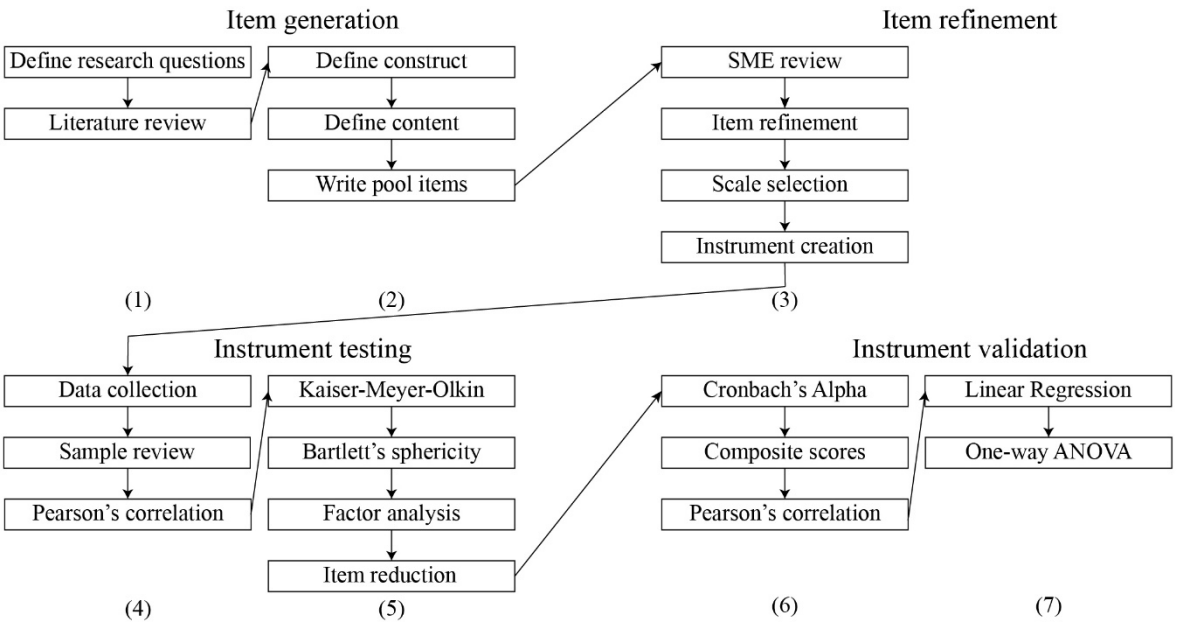

Figure 3.1.

SETED Methodology

Scale selection. Once a final pool of items was available, the next step was to apply a scale. An eleven-point Liker-type scale was used (Bandura, 2006; Carberry et al, 2010) ranging 0-10 in one-point increments. Bandura (2006) and Carberry et. al. (2010) both used three identifying statements along with the Likert-type scale. These identifying statements were located at the central point and the two outermost points. Bandura (2006) used descriptors worded "cannot do at all", "moderately certain can do", and "highly certain can do".

The instrument was created by combining the items developed in the previous step with the Likert-type scale described in this step. The items from the previous step are inclusive items, written with an included stem. This removed the need to place the stem in the instructions, as it was included with each item. It also allowed for the items to be randomly arranged as they did not need to remain together in order to keep the appropriate stem. An online software tool, Qualtrics, was used in order to host the instrument.

Instrument creation. The formatting of the instrument started with acknowledgement of informed consent, before progressing into the items themselves. The informed consent form can be found as Appendix I. After providing informed consent, the participant was taken to the items. If the participant declined to provide informed consent, the participant was taken to the exit page. After completing the items, the participant was taken to a page for the collection of demographic data.

The items were grouped into six item sections. The items were randomly arranged when placed on the instrument, but they remained consistent for all participants. There was one exception to the random item placement, the final item explicitly asked about teaching engineering design. The items were grouped into sections containing six items or less to help ensure that the identifiers for the Likert points were easily visible when responding to each item. The directions for each section were:

"Several statements of teaching tasks are described below that are performed in a classroom. These statements are related to your belief or feeling that you can successfully complete the task. Please rate how certain you are that you can complete each individual task."

Carberry et al (2010) and Yoon et al (2012) both gathered demographic information. Most demographic information is used to simply report the makeup of the sample population. Some demographic information, such as gender, was used to verify that no differences exist across that identifier. Teaching experience was demographic information gathered that finds direct use in the study. Following the examples set forth by the directing studies, the demographic data for this study was gender, ethnicity, age, years teaching, and years teaching engineering design. Demographic data was collected at the end of the instrument, rather than before, in the hopes that this measure reduced response bias. Given a demographic data question

that specifically referred to engineering design, this would have provided a strong indicator as to what is being researched. This consideration was given based upon a similar recommendation provided by Bandura (2006) in the *Guide to Constructing Self-Efficacy Scales*.

Bandura (2006) discussed methods for reducing response bias, including "if the scale is labeled, use a nondescript title such as "*Appraisal Inventory"* rather than *Self-efficacy*." This thought remained in consideration throughout the development of the instrument, as well as recruitment materials. Care was taken to minimize or eliminate the use of self-efficacy or engineering design in the wording of the instructions, recruitment materials, and the instrument itself.

Instrument testing

Instrument testing was the third phase of the development process. During this phase, data was collected and analyzed to determine next steps. This phase ended with factor analysis, which sought to indentify the underlying relationships between the items. If factor analysis had revealed issues with the items did not align with the theoretical underpinnings and required further refinement, the item refinement and instrument testing steps would have to be repeated until no such issues existed. This phase was broken into two steps: data collection and factor analysis.

Data collection. Carberry et al (2010) and Yoon et al (2012) both used an online survey hosting site for their instrument. Their recruitment efforts were also performed using email listings available to the researchers. This study followed both of these studies by sending out recruitment emails that provided a link to the instrument. The contents of this email can be found in Appendix I. Email listservs of professional organizations for engineering and technology educators were used for recruitment. The recruited participants were reflective of the intended

population for the instrument. The participants of this study were in-service teachers that teach engineering or engineering design in their classrooms.

Participants. The SETED sought to look at the self-efficacy of teaching engineering design. The population of interest was teachers that teach engineering design as part of their classes. In-service teachers were used in order to maintain the similarity to the Carberry et al (2010) study. This study recruited participants with varying levels of experience within the domain. The domain for this study was teaching engineering design. By recruiting in-service teachers this allowed for a wider range of experience levels in order to subdivide the sample into smaller groups. Several different email listings were used in order to increase the chances of having three stratified groups of experience like in Carberry, et. al. (2010). An explanation of how these groups were stratified is given during the analysis section of this methodology, later in the chapter. The use of email recruitment also hoped to enable a more diverse sample across gender, ethnicity, and regions. This was done in an attempt to reduce any result bias from a narrower sample focused on a singular location. Institutional review board (IRB) approval was sought for research with this population. The assessments and communications used in this study were those approved by the IRB.

Sample size. In order to complete the statistics outlined later in this chapter, a minimum of 50 participants were needed. The study intended on breaking the participants into three groups based upon responses to the demographic question about years of experience in teaching engineering design. The goal was to recruit at least 50 participants in each group in order to increase the strength of the statistical analyses performed. Both guiding studies (Carberry et. al., 2010; Yoon et. al., 2012) had to discard responses for a variety of reasons, so the assumption is this study will also need to do so. This increased the goal for a minimum number of participants.

This study increased the minimum to sixty as a safeguard against discarding results. Therefore, the goal for this study was to recruit at least two hundred participants, with at least fifty in each experience group, although the minimum of sixty total participants remains.

Data. The results were collected by the software tool, Qualtrics. The responses were kept there until the data collection process is complete. After the completion of the data collection process, it was downloaded so that it could be imported into a statistical software package Statistical Package for the Social Sciences, SPSS. Identifying information for participants was not downloaded, but the data storage still followed all appropriate guidelines set forth by North Carolina State University to maintain the confidentiality and privacy of the data and participants.

Sample review. Before an analysis could begin a review of the data and sample had to be completed. This review was intended to look for any responses that were deemed invalid and thus could have potentially skewed the results of the analysis. The first review of the data set looked for incomplete responses. Incomplete responses would have skewed the analysis by increasing the sample size without providing the corresponding data. The second review of data looked for any respondents that had no deviation in their responses. An average score of all items was found, an average score that resulted in a whole number was reviewed for variations in responses. After identifying which cases had no variation in responses, the time to complete the instrument was reviewed. A possible reason for no variation in responses was that the respondent simply clicked the same answers for each question in order to complete the instrument quickly. This type of response would also skew the analyses, as it would not appropriately reflect the participant's true answer to each item. One response showed no deviation and was completed quickly, upon reviewing the response it was found that the respondent had a large amount of experience. Given the experience of the respondent it was found that the responses followed

along with the theoretical underpinnings and most likely a valid response. Therefore, the second review of the sample identified any respondent that had no variation in responses. Once all of the invalid responses were identified, they were discarded from the sample before analyses were completed.

Carberry et. al. (2010) split their sample into three groups, current learners, little to no experience, and experienced. To follow along with Carberry et. al. (2010) this study would have needed to recruit pre-service teachers to have current learners. This did not fit with the population of interest, so the sample was split into three experience groups. Following the literature review, the new group of teachers was identified as zero to four years of experience. Following the literature review experienced teachers were nine plus years of teaching experience. This left those with six to eight years of experience. As this group fell between the new and experienced groups, they were identified as the intermediate experience group.

Item Correlation. Items were reviewed for multicollinearity (Yoon and Evans 2012, Yoon et. al., 2014), correlation with a value of greater than .85. This indicated that one, or more, items are measuring the same thing. The literature review presented three options for correlation, Pearson's (Carberry et al, 2010; Tasi, et al, 2021), Spearman's (Aydeniz et al, Malandrakis et. Al, 2019), and Polychloric (Sun and Rogers, 2020; Yoon, et al, 2014). This showed that testing for correlation was important, but the exact test was a little less clear. Given the balanced nature of reported use there was no immediately clear correct path. Each of these correlations made assumptions about the data. Both Pearson's and Polychloric carry with them the assumption of a normal distribution. Pearson's also assumed that the data is part of a continuous scale (Laerd, 2018e), while Polychloric makes no such assumption (Sun and Rogers, 2020; Yoon et al, 2014). Likert-type scales are by nature ordinal and not continuous, eliminating Pearson's as an option.

This left Polychloric and Spearman's correlation as options. The differing assumption was that Spearman's did not require the assumption of normality where as Polychloric did. A sufficiently large sample would have allowed use of the central limit theorem. Due to the underlying assumptions of the analyses, Spearman's correlation was performed as it best fit the ordinal data and did not require an assumption of normality. The guiding study (Carberry et al., 2010) utilized Pearson's correlation, to be consistent with the guiding study Pearson's correlation between each item was still performed. These values can be found in Appendix B for Pearson's and Appendix C for Spearman's.

KMO. Kaiser-Meyer-Olkin measure of sampling adequacy was an analyses performed to determine if the sample was viable for factor analysis (Aydeniz, et al., 2020; Kukul and Karatas, 2019; Malandrakis, et al., 2019; Sun and Rogers, 2020; Tsai, et. al., 2018; Tsai, et. al., 2021; Xu, et al, 2019). Possible values ranged from 0.0 to 1.0 with a value greater than 0.50 being indicating factor analysis would be appropriate for the data.

Bartlett. Bartlett's test of sphericity was the second test used to determine the data's eligibility for factor analysis (Aydeniz, et. al., 2020; Kukul and Karatas, 2019; Malandrakis, et al., 2019; Sun and Rogers, 2020; Tsai, et. al., 2018; Tsai, et. al., 2021; Xu, et al, 2019). This analysis tested the hypothesis that the correlation matrix was actually an identity matrix and thus unrelated. A statistically significant result, $p<0.05$, would indicate that the variables are related and suitable for factor analysis.

Factor analysis. After determining the data was appropriate for factor analysis, the next step of the instrument testing process was factor analysis itself. This study used exploratory factor analysis, EFA (Aydeniz et al., 2020; Kukul and Karatas, 2019; Malandrakis, et al., 2019; Sun and Rogers, 2020; Tsai, et. Al., 2021; Xu, et al, 2019; Yoon, et al., 2012; Yoon, et al., 2014).

In Yoon et. al. (2012) they indicated that their data did not follow a normal distribution. Their data was also ordered and could be treated as categorical, a result of using a Likert type scale. This study also used a Likert type scale, and it is expected that it will also not follow a normal distribution. This made a strong case for using EFA. The goal of EFA was to assist with instrument refinement through item retention and removal, as well as identifying the underlying factors of an instrument.

There were two major types of rotation within EFA, orthogonal and oblique. This study expected the underlying factors to be related, not independent because of the Yoon et. al. (2014) study. Therefore, an oblique rotation was used as it is the appropriate method for interrelated factors (Aydeniz et al., 2020; Kukul and Karatas, 2019; Malandrakis, et al., 2019; Sun and Rogers, 2020; Tsai, et. Al., 2021; Xu, et al, 2019; Yoon, et al., 2012; Yoon, et al., 2014). The results of the factor analysis needed to be further reviewed. If this occurred the item with the strongest factor loading was kept, and the others discarded. EFA then provided a number of underlying factors within the instrument. When determining how many of the underlying factors were appropriate, there were a few methods available. To keep in line with the guiding study of Yoon et. al. (2012), this study used Kaiser's (1960) method. This method suggested that only eigenvalues greater than one should be used (Aydeniz et al., 2020; Kukul and Karatas, 2019; Malandrakis, et al., 2019; Sun and Rogers, 2020; Tsai, et. Al., 2021; Xu, et al, 2019; Yoon, et al., 2012; Yoon, et al., 2014).

After determining the number of factors, the loading of each item onto a factor was determined. Each item should load on to only a single factor. Any items loading onto more than one factor were removed. The same was true for any items not loading on to any factors. Yoon et. al. (2012) used a cutoff of .40. This loading cutoff was dependent upon sample size, so had to

be determined after data collection. This value was found to be .370, based upon eighty-one respondents. At the end of loading, any factor that had no items were discarded (Yoon et. al., 2012). The result of this factor analysis was the identification of latent relationships between the items. These latent relationships should be consistent with the theoretical underpinnings for the instrument to be considered valid.

In addition to EFA, there was also confirmatory factor analysis, or CFA. CFA was typically used later in research once the underlying factors had been identified (Kukul and Karatas, 2019; Sun and Rogers, 2020; Xu, et al, 2019; Yoon, et al., 2014). CFA should also be used when there are an expected number of factors. Due to a single construct, this study expected only a single factor. Carberry et al (2010) also used CFA in order to confirm that each of the individual items were consistent across constructs. For instance, redesigning, was analyzed across all four self-concepts to ensure that the item itself was consistently represented in each self-concept. CFA was not performed as a part of this study.

Before continuing, the factor analysis was reviewed to confirm that it aligned with the theoretical underpinnings of the study. Factor analysis also indicated that three of the items needed to be reviewed and either reworded or removed before continuing. Upon review of the items and data it was determined that the items were unable to be reworded. As a result, it was decided that they should be removed from the instrument. The remaining items demonstrate similar concepts within their wording, so removal of any item does not result in the loss of that theoretical construct. In addition, the removal of the three items does not drop result in a factor having no items loaded so the number of factors remains the same.

Instrument validation

The fourth phase of this study was the validation phase. In this phase evidence to answer the research questions was directly sought, although evidence towards answering them may be provided from any of the phases. This phase was broken into two steps each with several sub-steps. The first step looked at reliability and breaking the instrument apart based upon the results of the factor analysis. The second step was performing statistical tests to provide evidence for validity of the instrument.

Reliability. Cronbach's α provided evidence of a common construct and was critical in making a case for validity (Aydeniz et al., 2020; Carberry et al 2010; Kukul and Karatas, 2019; Malandrakis, et al., 2019; Sun and Rogers, 2020; Xu, et al, 2019; Yoon, et al., 2012; Yoon, et al., 2014). It did this by determining the one-dimensional nature of the instrument. There were a range of values, with 0.70 to 0.95 being ideal (Tavakol, 2011; Aydeniz, et al, 2020). Any values lower indicated the instrument was unreliable. Higher values may have been a result of multiple items measuring the same thing, also an indicator of an unreliable instrument. By waiting until after factor analysis and any resulting item refinement, the chance of values that were too high was reduced. Yoon et al (2012) also suggested testing the removal of items in order to see the effect on Cronbach's α. Due to the mention of leaving all items in their study because removing them did not increase Cronbach's α, the implication of Yoon et al (2012) was that if the removal of an item increases the reliability of the instrument this was an acceptable measure.

Composite score generation. After reliability testing the next step was to create composite scores for each factor (Carberry et al 2010; Malandrakis, et al., 2019; Tsai, et. al., 2018; Tsai, et. al., 2021; Xu, et al, 2019; Yoon, et al., 2012; Yoon, et al., 2014). As a result, this analysis was dependent on the results of EFA. Carberry et al (2010) used an item that explicitly

states performing engineering design, the topic of their instrument. The SETED also contained a single item that related to teaching engineering design. Carberry et al (2010) abbreviated these explicitly stated items to ED, a tactic this study repeated. This was to differentiate it between the composite score along a factor, which was referenced to as EDP (Carberry et al, 2010). The EDP in the Carberry (2010) study was a composite score for the entire self-efficacy portion of the instrument. This study referred to the composite score of the instrument as engineering design composite, or EDC. During EFA this study found six underlying factors. As a result, each of these factors was named ED1-ED6 with the number corresponding to the factor number.

The goal of creating a composite score was to be able to use parametric analyses on the data. A simple mean would have generated a combined score for multiple items, but Glen (2015) stated "the mean in a Likert scale can't be found because you don't know the distance between the data items." This was a result of ordinal data, where the distance from one point to the next was not a fixed value. Starkweather (2012) noted that a simple mean was also not appropriate because "it treats each question as contributing to the composite score equally, which is often not the case when one considers the latent variable structure." Starkweather (2012) went on to suggest composite scores for Likert scales should be weighted means based upon the results of factor analysis, a common process (Malandrakis, et al, 2019; Tasi, et al, 2018; Tsai, et al, 2021; Xu, et al, 2019; Yoon et al, 2014). This study used this process in order to generate each of the ED1-6 scores as well as the EDC. The ED1 scores were created by taking the sum of the loaded items responses and multiplied by the items factor loading for the appropriate individual item. This sum was then divided by the sum of all the factor loadings for that factor, resulting in a weighted mean. The EDC was a composite score for the entire instrument, meaning it represented the combination of each factor. Given Starkweather's (2012) guidance for creating

composite score, the EDC was also calculated as a weighted mean. This calculation used each of the factor scores, ED1-6, and multiplied them by their respective variances explained. These values were summed and then divided by the sum of the variances explained. The variances explained was used because it represented the weight that each factor carried within the instrument.

Pearson's correlation. The correlation between each of the factor scores, the composite score, and the specific engineering design item was also tested (Aydeniz, et al., 2020; Carberry, et al, 2010; Malandrakis, et al, 2019; Sun and Rogers, 2020; Tsai et al, 2021; Yoon et al, 2014). Earlier in this chapter three correlation analyses were discussed and the assumptions of each noted. With the creation of a composite score, the data type was transformed to be continuous. The sample size was large enough to use the central limit theorem, meaning that Pearson's correlation became the appropriate analysis for the factor correlations. This study found the correlation between the EDC, each of the ED1-6 factors, and ED of the sample. Since the EDC was composed of the ED1-6 factors it was expected to have a strong positive correlation. Statistically significant correlations would indicate that all the factors were related. This would provide further evidence that all items measured the same construct.

ANOVA versus linear regression. After generating composite scores the literature review showed that further analysis was performed on the result. The goal of these analyses was to demonstrate predictive validity by showing alignment to the self-efficacy theory, which used experience as an independent variable and the composite score(s) as the dependent variable. The three options found were the paired t-test (Malandrakis, et al., 2019; Tsai, et. Al., 2018), the one-way ANOVA (Carberry et al 2010; Tsai, et. Al., 2018), and regression (Malandrakis, et al., 2019; Tsai, et al., 2021). The ANOVA was slightly more common and the guiding study,

Carberry et al (2010) used a one-way ANOVA. Laerd (2018d) lists the following 6 assumptions of a one-way ANOVA test:

> "Assumption #1: Your dependent variable should be measured at the interval or ratio level.
>
> Assumption #2: Your independent variable should consist of two or more categorical, independent groups.
>
> Assumption #3: You should have independence of observations.
>
> Assumption #4: There should be no significant outliers.
>
> Assumption #5: Your dependent variable should be approximately normally distributed for each category of the independent variable.
>
> Assumption #6: There needs to be homogeneity of variance."

By nature, Likert data is ordinal. This means that the first assumption was violated by the use of Likert items. Glen, S. (2015) states "for a series of individual questions with Likert responses, treat the data as ordinal variables. For a series of Likert questions that together describe a single construct, treat the data as interval variables." With this guidance a composite score, which represented a single factor, was treated as an interval variable. This allowed for the first assumption to be met. The grouping of respondents based upon teaching experience that did not allow for overlap answered assumption two. Assumption three was passed with each respondent being separate. A review of the data also showed that assumption four was met. Assumption five was the assumption of normality. While the central limit theorem allows for large sample sizes to be treated as normal, some of the groups in this study were not large. This means that in order for the ANOVA to be an acceptable test for the ED1-6 factors and EDC, a test for normality would be required. If the samples are found to be normal, then the ANOVA's assumptions are met and

it remains a valid analysis. If the normality assumptions were not met, then the ANOVA would not be the correct analysis. However, for the ED item that was not transformed from an ordinal variable, non-parametric tests are required. None of the surveyed literature indicated the use of non-parametric tests, with the Kruskal-Wallis test being the appropriate test for ordinal data. Laerd (2018b) identified the following six assumptions for linear regression:

> "Assumption #1: Your two variables should be measured at the continuous level.
>
> Assumption #2: There needs to be a linear relationship between the two variables.
>
> Assumption #3: There should be no significant outliers.
>
> Assumption #4: You should have an independence of observations.
>
> Assumption #5: Your data needs to show homoscedasticity
>
> Assumption #6: You need to check that the residuals (errors) of the regression line are approximately normally distributed".

The assumptions of linear regression were a better fit for the data present. Likert data was still ordinal, however the composite score generation outlined in the discussion of the ANOVA test allowed for the transformation of the data to a non-ordinal nature. The first assumption also represented a better fit to the independent variables gathered on the instrument, experience was measured on a continuous level by years. The second assumption could be answered by viewing a scatterplot (Laerd, 2018b) as well as through correlation. There were no significant outliers found, meeting the third assumption. The observations were independent meeting the fourth assumption. The fifth and sixth assumptions could only be verified after the linear regression analysis was performed; although Laerd (2018b) noted that verifying the fifth assumption with real-world data is difficult.

Paired T-test. During the discussion on selecting the correct statistical analysis it was noted that the ANOVA and linear regression were the more appropriate tests. However, there is one use of the paired t-test that remains appropriate, this instrument included an item specifically asking about the self-efficacy of teaching engineering design. This item acts as a control, if the EDC represented the self-efficacy of teaching engineering design, then it would have been expected that a paired T-test between each respondents EDC and ED score would yield no differences (Carberry et al, 2010; Malandrakis, et al, 2019; Tsai, et al, 2018). However, the control item remained untransformed and thus was ordinal in nature. As a result, the assumptions of the paired t-test were not met. Instead, the nonparametric Wilcoxon signed-rank test was used.

ANOVA. A one-way ANOVA was performed to look for differences based upon gender. The goal of this test is to not reject the null hypothesis and show that the composite scores of the instrument are not affected by gender. (Carberry et al, 2010; Yoon et al, 2014; Malaandrakis, et al, 2019, Tsai et al, 2021).

Summary

This study used a four phase, multistep process informed by the literature review and aligned with the guiding study, Carberry et al, 2010. The generation of research questions resulted in a review of literature to find an answer to these questions. Finding no answer, this study looked to use an established instrument to answer the original questions. Finding no appropriate instrument, the scope of the study changed to developing an instrument that could be used to answer the original questions. With this change in scope, new research questions were created. These questions focused on the creation and validation of an instrument to test the self-efficacy of teaching engineering design. A second review of literature was used to guide the instrument development process, with a guiding study being selected to set the framework on

which the study was built. This framework was supported based upon the findings in the literature.

The second literature review also focused on defining the construct of self-efficacy and the content domain of engineering design. These findings were presented in the previous chapter. Once the literature review was completed, initial items were developed by the researcher. These items were composed of one of two self-efficacy stems and a teaching engineering design content item. This created an initial item pool. The initial item pool was reviewed by SMEs for face validity and edits suggested. Edits were made to the items in a cyclical review process until no further edits were recommended. Once the item pool was completed, the items were given a scale and the instrument created.

After the instrument was created and hosted online, participants were recruited. At the close of the recruitment time frame, the data was reviewed and filtered based upon selection criteria. This generated a final data set which was then analyzed for correlation of items and appropriateness of factor analysis. Finding factor analysis to be appropriate, EFA was performed and six latent factors discovered. A forced factor EFA confirmed the presence of a single construct. As a result of factor analysis, it was suggested that several items be removed.

Reliability testing was then completed and composite scores for each factor and the instrument as a whole created. These composite scores were then analyzed using linear regression to find if experience had an effect on the scores; a one-way ANOVA confirmed that there were no difference in scores based upon gender.

CHAPTER 4: RESULTS

Introduction

The purpose of this study was to design and validate an instrument to test the self-efficacy of teaching engineering design for in-service teachers in Technology and Engineering education. There were one hundred responses to this study, and the results were generated from eighty-one of these responses. The results were presented in the order in which the analyses were performed, following along with the methodology.

Sample review

There were one hundred respondents to this study. Sixteen respondents did not make it to the end of the instrument. One respondent made it to the end but did not provide any data. One response was finished but only contained one answer. One response was finished but contained two missing data points. After removing incomplete responses, the data was reduced down to eighty-one responses. Eight responses had a whole number average, but only six showed no variation. These six cases were reviewed for completion time. The average time to complete the instrument was five minutes one second with a median of seven minutes and fifty-three seconds. Only one case fell significantly outside of the average range. This response fell in line with the theoretical underpinnings of the study, so was determined to be valid. The result is that the initial one hundred responses parsed down to a final set of eighty-one valid responses.

Correlations

Following along with the guiding studies the first step was to find Pearson's correlation of each item, reviewing the correlations for values greater than 0.850 and less than 0.200. Table 4-1

below shows a sample of the results of the Pearson's correlation, with the full results being contained in Appendix B.

Table 4-1.

Example Pearson's correlations

Variable	1	2	3	4	5	6
1	1.00					
2	.497**	1.00				
3	.590**	.670**	1.00			
4	.566**	.351**	.482**	1.00		
5	.247*	.103	.275*	.462**	1.00	
6	.230*	.002	.318**	.374**	.484**	1.00

*p<.05. **p<.01. ***p<.001

The guiding study performed Pearson's correlation, although the data from Likert items was ordinal. The methodology outlined in the previous chapter identified Spearman's correlation as a more appropriate test. As a result, Spearman's correlations were also found. A sample of the results can be found in Table 4-2, with the full results being located in Appendix C.

Table 4-2.

Example Spearman's correlations

Variable	1	2	3	4	5	6
1	1.000					
2	.542**	1.000				
3	.564**	.694**	1.000			
4	.626**	.465**	.531**	1.000		
5	.412**	.255*	.355**	.559**	1.000	
6	.307**	.096	.361**	.491**	.529**	1.000

*p<.05. **p<.01. ***p<.001

Both results were reviewed for multicollinearity and found none. Any such results would have been flagged for further review after the factor analysis.

KMO

After establishing the data set, according to the methodology the first step was to test for appropriateness of the data for factor analysis. Kaiser-Meyer-Olkin measure of sampling adequacy, KMO, and Bartlett's test of sphericity were performed. KMO values range between zero and one, with values closer to one being desired and a threshold of .6 being required before continuing with factor analysis. This study found KMO to be .826 indicating the data was adequate for performing exploratory factor analysis.

Bartlett's

After determining the KMO indicated the data was satisfactory for factor analysis, Bartlett's test of sphericity was performed. Bartlett's [X^2=1782.995, 378, p<.001] indicated that the data was not an identity matrix and thus factor analysis would be appropriate.

Factor Analysis

Exploratory factor analysis was performed using IBM Statistical Package for Social Sciences Statistics 26, SPSS. Exploratory factor analysis was performed using principal axis factoring and a Promax rotation with Kaiser Normalization of four. Promax is a form of oblique rotation, which should be used for items that are not independent. This analysis found six factors that explained a total of seventy-three percent of the variance, using Eigenvalues greater than one as the cutoff.

When reviewing items and factor loadings, Thurstone's (1947) criteria for a simple structure is commonly used to determine which loadings are considered significant and salient. Brown (2009) provides some guidance on determining cutoff scores that are significant based upon sample size. With a sample below the discussed one hundred, the cutoff of .300 is not

appropriate. However, with a sample of eighty-one the higher value of .400 could be used for salient, if not significant loading.

Table 4-3.

Example factor loadings

	Factor Loading					
Item	1	2	3	4	5	6
1	0.243	0.068	**0.554**	0.091	0.218	-0.244
2	-0.219	-0.077	**0.908**	0.185	-0.084	0.09
3	0.175	-0.028	**0.729**	-0.045	0.075	0.277
4	0.05	0.068	0.302	-0.044	**0.621**	-0.133
5	-0.088	0.079	-0.007	-0.13	**0.696**	0.126
6	**0.450**	-0.18	-0.188	-0.004	**0.552**	0.297

Table 4-3, shows an example of the factor loads found with bold values indicating a value greater than .400. Using this score of .400 as a cutoff for factor twenty-four items load onto only a single factor. Items six and eighteen loaded onto two factors. Items nineteen and twenty-three failed to load onto any of the factors. As such, all four items were removed. Appendix D shows the full factor pattern matrix. Appendix E displays items sorted factors and provides their respective loadings.

Reliability

Cronbach's α was calculated for the remaining twenty-four items and found to be 0.947. This level indicates that the instrument has a high internal reliability. Following the guidance of Yoon, et. al. (2012), Cronbach's α was computed for the instrument with the removal of each individual item. Cronbach's α remained the same or decreased with the removal of each item, except for item five. With the removal of item five, Cronbach's α increased to 0.948. Since the removal does not result in Cronbach's α increasing above the upper bounds, this implies that item five should be considered for removal. This case is strengthened by item five having the lowest

loading during the forced factor EFA. The full listing of Cronbach's α can be found in Appendix F. The base score of all twenty-four items is provided first. Afterwards the new value with the indicated item removed is provided.

Validation

In establishing a case for validity several analyses were performed. The results of each can be found below. They have been divided into sections for ease of reporting. Before performing these analysis, composite scores were generated for each of the six factors found during the exploratory factor analysis. Each of these were identified as ED1-ED6 to correspond with each factor. These composite scores were then used to make a holistic composite for the entire instrument, labeled the EDC. There was a single item that remained labeled as ED and directly stated the question of "I believe I can teach students engineering design".

Pearson's factor correlations

When Pearson's correlation was run between the ED and EDC, they were found to have a significant correlation $r(80)=.62$, $p<.001$. Following along with the analysis of the ED and EDC Pearson's correlation was found between each factor score, ED, and the EDC. Table 4-4 shows the results of this analysis.

Table 4-4.

Pearson's correlation of factors

Factor	ED1	ED2	ED3	ED4	ED5	ED6	ED (Item)
ED1							
ED2	.689**						
ED3	.470**	.598**					
ED4	.484**	.523**	.431**				
ED5	.359**	.335**	.417**	.365**			
ED6	.432**	.438**	.436**	.370**	.380**		
ED	.666**	.420**	.315**	.483**	.347**	.228*	
EDC	.967**	.788**	.621**	.599**	.487**	.561**	.655**

*p<.05. **p<.01. ***p<.001

The first factor, ED1, was found to have strong positive correlations with ED2 [$r(80)=.728$, $p<.001$], ED3 [$r(80)=.527$, $p<.001$], ED4 [$r(80)=.502$, $p<.001$], ED5 [$r(80)=.382$, $p<.001$], ED6 [$r(80)=.453$, $p<.001$], EDC [$r(80)=.866$, $p<.001$], and ED [$r(80)=.664$, $p<.001$]. ED2 also had strong correlations with the other factors, ED3[$r(80)=.647$, $p<.001$], ED4 [$r(80)=.568$, $p<.001$], ED5 [$r(80)=.374$, $p<.001$], ED6 [$r(80)=.478$, $p<.001$], EDC [$r(80)=.910$, $p<.001$], and ED[$r(80)=.480$, $p<.001$]. ED3 continued with this trend, correlating with ED4 [$r(80)=.482$, $p<.001$], ED5 [$r(80)=.471$, $p<.001$], ED6 [$r(80)=.475$, $p<.001$], EDC [$r(80)=.770$, $p<.001$], and ED [$r(80)=.379$, $p<.001$] positively. ED4 also demonstrated positive correlations with the other items ED5 [$r(80)=.443$, $p<.001$], ED6 [$r(80)=.437$, $p<.001$], EDC [$r(80)=.715$, $p<.001$], and ED [$r(80)=.526$, $p<.001$]. ED5 also had positive correlations with the remaining items ED6 [$r(80)=.404$, $p<.001$], EDC [$r(80)=.566$, $p<.001$], and ED [$r(80)=.389$, $p<.001$]. ED6 also positively correlated with the EDC [$r(80)=.609$, $p<.001$]. ED6 and ED [$r(80)=.264$, $p<.05$] had a less significant correlation, but a positive one none the less. The full set of correlations are listed in Appendix G.

Regression models

According to the self-efficacy theory on which this study was based, it was expected that experience teaching engineering design would have a positive effect on the composite score. This study also collected information on general teaching experience. Several different models of linear regression were performed to search for the best model. The first model found there was a significant effect between SETED composite score and experience teaching engineering design ($F(1,79)=8.232$, $p=.005$, $R^2=.083$). Teaching experience alone did not have a significant effect on the SETED composite score ($F(1,79)=.194$, $p=.661$, $R^2=.002$). However, both experience teaching engineering design and experience teaching engineering design did have a significant effect on the SETED composite score ($F(2,78)=6.469$, $p=.003$, $R^2=.142$). Further investigation of the individual predictors found that both experience teaching engineering design ($t=3.566$, $p=.001$) and general teaching experience ($t=-2.087$, $p=.040$) were significant predictors at the $p<.05$ level although only experience teaching engineering design remained significant at the $P<.01$ level.

Wilcoxon signed-rank test

The Wilcoxon signed-rank test found that there was a significant difference ($Z=-3.885$, $p<.001$) between the composite score (M=8.5607, min=5.46, max=10) and engineering design control item (M=8.88, min=2, max=10). There were 56 negative rank cases, in which the composite score was lower than the control item. There were 24 positive rank cases and 1 tie.

Demographics

The gathered demographic data indicates three respondents teach at the 6-8 grade level, thirty-one respondents teach at the 9-12 grade level, forty-eight respondents teach at the 12+ grade level. Seventy-nine respondents were from the United States, one respondent was from

Canada, and another from Latvia. Fifty-seven of the respondents were male and twenty-four females. One respondent identified as Asian, seventy-nine respondents White, and one other. Table 4-5 One-way ANOVA gender results, shows no significant differences on any of the factor scores or the composite score based upon gender.

Table 4-5.

One-way ANOVA gender results

Factor	df	F	Sig
ED1	1	0.138	0.711
ED2	1	0.360	0.550
ED3	1	0.928	0.338
ED4	1	0.452	0.504
ED5	1	0.103	0.749
ED6	1	0.797	0.375
EDC	1	0.000	1.000

The sample population did have vastly different sample sizes based upon gender. While different sample sizes in themselves are not an issue, similar sample sizes are required in order to bypass the homogeneity of variance assumption within the one-way ANOVA. As a result of the sample size issue, Levene's test was performed to identify if this assumption held true or was violated. Table 4-6 Levene's test gender results, shows the results of these analyses.

Table 4-6.

Levene's test gender results

Factor	df	F	Sig
ED1	1, 79	0.138	0.711
ED2	1, 79	0.360	0.550
ED3	1, 79	0.928	0.338
ED4	1, 79	0.452	0.504
ED5	1, 79	0.103	0.749
ED6	1, 79	0.797	0.375
EDC	1, 79	0.000	1.000

Summary

This chapter presented the results of this study beginning with the data collection refining the sample down from one hundred to eighty-one based upon exclusion criteria. These eighty-one respondents were then used to find the correlation of the items to ensure the values were between 0.850 and 0.200. The KMO was found to be 0.826 indicating EFA could be performed. After correlation of the items, exploratory factor analysis found six underlying factors, with three items loading onto multiple factors. These items were excluded and composite scores for each of the six factors and the instrument as a whole were created. Cronbach's α for all items was found to be 0.948. Regression models showed that general teaching experience was not a predictor of composite SETED score, while experience teaching engineering design was. Regression models showed that the combination of experience teaching engineering design and general teaching experience was a stronger model than simply experience teaching engineering design alone.

Chapter 5: DISCUSSION

Introduction

The purpose of this study was to design and begin to validate an instrument to test the self-efficacy of teaching engineering design. This goal resulted in four research questions being presented:

Q1. Is there evidence that the SETED scale is reliable?

Q2. Is there evidence of validity of the SETED scale in the theory of self-efficacy?

Q3. Is there evidence that the SETED items represent the domain of teaching engineering design?

Q4. Is there evidence that the SETED scale is a valid instrument?

The answer to these questions lies within the results presented in the previous chapter. Additionally, previous chapters discussed the AERA (2014) *Standards for education and psychological testing* and identified standards relevant to this study. Each of these standards relate to the research questions, but they will be discussed together as a unit.

RQ 1: Reliability

The first question of this study was: does the SETED show evidence of reliability? To answer this question the first step involved reviewing the inter-item correlations. This check was done before factor analysis. In reviewing the significant correlations between each of the items it was found that all items had a significant correlation with at least one other item between the values of .200 and .850 (Yoon et al, 2014; Carberry et al, 2010, Tasi et al, 2021; Sun and Rogers, 2020). This positive indicator allowed progression forward, as all items were purported to measure the same construct any item that had no significant correlations or a significant negative

correlation would indicate a deviation from the theoretical underpinnings. A review of the correlations found no significant, negative correlations; however, a significant result of less than .200 also indicated that an item only had very loose associations with the other items which would have also suggested that an item was not measuring the same construct as the other items. A review of the correlations found no items with a significant result below .200. While the lower bound of significance was .200, there did exist an upper bound as well. Significant values greater than .850 indicate multicollinearity (Yoon et al, 2014) or that two, or more, individual items likely measure the same thing. A goal of instrument development is efficient measurement of the construct, meaning duplicate or redundant items unnecessarily increase the length of the measurement. Items eight and nine had the highest significant correlation with a value of .818, within the acceptable range. The instrument passes the first check for reliability by demonstrating that all items are related but none are duplicated measures. The second check was Cronbach's α (Aydeniz et al, 2020; Kukul and Karatas, 2019; Malandrakis, et al, 2019; Sun and Rogers, 2020).

In the previous section Cronbach's α was found to be 0.947. Removal of item five increases Cronbach's α to 0.948. Both of these values are within the upper bounds of 0.950 for reliability. This means that in either form, with or without item five, the instrument has provided evidence for reliability. This reliability check was performed after factor analysis, which removed several of the items from the instrument. As removal of item five did not greatly increase Cronbach's α (Yoon et al, 2014), and it was found to significantly load onto a factor, it was left in the instrument. This data is also intended to answer AERA (2014) standard 2.3 with the reporting of reliability measures for each composite scores and total scores, in this case simply a total score.

RQ 2: Self-efficacy validity

The second question was: does the SETED scale provide evidence for validity in the theory of self-efficacy? In order to provide evidence of validity, it would be expected that the results of the instrument follow along with self-efficacy theory. This means that as experience in teaching engineering design increases, the corresponding composite score should increase (Carberry et al, 2010). It also means that each of the items should be written in such a way that they appear to be speaking to the respondent's belief in their ability to complete a task (Bandura, 2006). The answer to this question has two parts. The first consideration comes in the form of face validity. As a part of item creation, SMEs were used in a cyclical process for item refinement (Kukul and Karatas, 2019; Tsai et. Al, 2021; Xu, et al, 2019). The SME approval, who were recruited for their knowledge of engineering design and self-efficacy, indicates that the wording of each item appeared to represent the construct of self-efficacy. This passed the first test of validity for self-efficacy.

The second test was if the results aligned as expected with the theory. The linear regression model (Malandrakis, et al., 2019; Tsai, et al., 2021) showed predictive validity by indicating that the engineering design composite, EDC, score $F(2,78)=6.469$, $p=.003$, $R^2=.142$) showed a positive association with experience both in teaching engineering design and in general. Experience teaching engineering design ($t=3.566$, $p=.001$) and general teaching experience ($t=-2.087$, $p=.040$) both had significant associations. However, experience teaching engineering design had results significant at a higher threshold and was a positive predictor, which matches the self-efficacy theory. A simplified linear regression model using only teaching engineering design as the independent variable still showed significant association ($F(1,79)=8.232$, $p=.005$, $R^2=.083$) although the results had a lower percentage of variance

explained. This model represents the best case for validity as it aligns with the fundamental theory of self-efficacy increasing with mastery experience. The model using only teaching experience did not show a significant association with the composite score. This data shows that experience teaching engineering design is the largest factor associated with the SETED composite score, following in line with the self-efficacy theory of greater experience having a positive association with the composite score. The fact that general teaching experience did not have a significant association alone further aligns with self-efficacy's domain specific theory (Zee and Koomen, 2016). While the simple model involving only experience teaching engineering design is the best fit to the theoretical underpinnings and has a significant, positive result, this study recommends keeping the model that uses both sets of experiences. The reason for this is discussed later in this chapter.

RQ 3: Engineering design validity

The third question proposed by this study was: does the instrument, SETED, represent the domain of teaching engineering design? Like the self-efficacy theory of question two, this question is answered through multiple data points. The first data point is again SME validation (Kukul and Karatas, 2019; Tasi et al., 2021; Xu, et al, 2019).. SMEs with experience teaching engineering design were recruited to review all of the items and provide feedback. This feedback was used to refine the items until no further feedback was provided. With the final set of items being approved by the SMEs, the items were deemed to have face validity within the teaching engineering design domain. This face validity provides one positive indicator, but the case for validity would be stronger with a statistical analysis to back it up. For the question of teaching engineering design, factor analysis was used to gather evidence for validity.

Factor analysis

Factor analysis found six underlying factors in this study. When creating a new instrument many studies, such as the guiding studies of Carberry et. Al. (2010) and Yoon et. Al. (2012), run multiple constructs within their instrument. In these cases, factor analysis shows that the items remain consistent and load onto a singular construct and that it is the intended construct. With no pre-existing scale available at the time, this study did not run a parallel instrument or construct. As a result, it should be expected that factor analysis did not reveal a single factor of teaching engineering design. Factor analysis seeks to reduce the observed variables into a smaller number of latent variables. This study obtaining multiple factors in itself is not problematic. The items were created under the premise of engineering design being a complex topic that itself had several underlying factors. Through a literature review this study identified the characteristics of good engineering design problems (Lammi, 2011; Lammi et. Al, 2018). Items were then created based upon characteristics. Factor analysis seeks to discover underlying, latent, variables (UCLA 2). As a result, if the items within each factor displayed a theme, this was consistent with the theoretical underpinnings. The guiding literature for understanding engineering design identified only three themes, but six factors were found. This deviates from the theoretical underpinnings, which could be problematic. However, upon looking at each of the factors the results did not deviate from the theory as far as the simple comparison of themes to factors indicated.

Factor six contains a single item, creating metrics to assess student success. This is the only item on the instrument that relates directly to a teaching task, but not specifically a characteristic of engineering design. This item is a characteristic of teaching pedagogy, resulting

in the preliminary name of pedagogy. This aligns with the theoretical underpinnings that identified pedagogy as its own theme.

This leaves five more factors to review and verify alignment with the theoretical underpinnings. They will continue to be looked at in a reverse variance order, meaning the factors explaining the least amount of variance will be thematically reviewed first. Factor five contained three items. Each of these items express different methods of visually representing a design concept. There is only one other item that relates to a visual representation of ideas, item seven. There does exist a clear difference between item seven and the other three items in factor five. The three items in factor five are all advanced forms of modeling, which resulted in this factor being named advanced modeling. They represent creating a physical model or a virtual model. This virtual model is broad, being defined only as a simulation. The final item was removed but indicated visual models of a design. It was removed for loading onto two items, but it loaded more strongly onto factor five rather than one. This removed item still continued with the theme of advanced modeling. The literature review also indicated that modeling, which could be physical or a simulation (Lammi, et al, 2018) was one of the two hallmark artifacts of engineering design. As a result, factor five did align with the theoretical underpinnings of this study.

Factor four has four items. At first glance these items represent a larger gambit of concepts including communicating ideas, design teams, iterative design, and dealing with failure. McFadden and Roehrig (2019) identified the first three items as important in their study noting that engineers did not work in isolation and must be able to communicate to work effectively in design teams, and that engineering design is a naturally iterative process. The only item not covered by them was dealing with failure. Wendell et al (2017) noted that an instructor's ability

to help students cope with failure and recognize it as part of the design process was a distinguishing characteristic of an experienced teacher. When working on a multidisciplinary team, this communication may appear as more than just visual and include written and oral communication. All of these concepts, including those removed, aligned with hallmarks of experienced engineers and skills that are developing or missing from novice engineers. As a result, this factor was named experienced engineer skills. These are the advanced engineering design skills, which are more difficult to impart or imparted latter in the process of teaching engineering design, once students have mastered the basics. This aligned with the theoretical underpinnings of this study.

Factor three had four items: multidisciplinary problems, systems thinking, open-ended problems, and thinking mathematically. Each of these items represents Lammi et al (2018)'s theme of challenge attributes which identified "relevance, open-endedness, systems perspective, and transdisciplinarity" as the main components of their "challenge attributes" theme. In fact, each of these items directly states one of these themes except for relevance and thinking mathematically being a mismatch. However, mathematics is included by Lammi et al (2018) in their definition of transdisciplinarity stating "engineering is not just applied math" showing that they regarded math as a discipline in and of itself. Each of the three remaining concepts all related to the complexity of the design challenge. As a result this factor was named complex engineering design problems. An alternate name for this factor is engineering design problem attributes, which aligns with the theoretical underpinnings. However, this name was not used due to a concern with creating confusion about these being the only attributes of engineering design challenges. This was the second factor to match with the theme structure used for item generation.

Factor two has ten items. A review of these items shows a very clear theme, working with and defining the problem. This factor has been named problem identification to reflect this. McFadden and Roehrig (2019) noted that engineering design is a cyclical process, and Wendell et al (2017) noted the cyclical nature as well, but added that experienced engineers spend more time within this process defining the problem and systematically evaluating the designs. With an iterative process, the definition of the problem would be continually updated, requiring purposeful questioning to correctly define the problem. These questions and time spent defining the problem also help deal with ambiguity. Chao et al (2017) noted that even through trial and error learners may stumble upon a correct answer, but in order to understand that they have a correct answer the learner would have needed to frame the problem in such a way to recognize that their happenstance has led them to a solution. The items within this factor all represent problem definition, a skill which is supported by the theoretical underpinnings of this study. The variety of items indicates the complexity that this skill represents.

Factor one has seven items. A first review of these items showed a wide variety items that had no consistent theme. It was important at this point to remember the foundation on which this study was based, that engineering design was more complex than the typically simple representation of it that was presented in educational systems. Carberry et al (2011) and Wendell et al (2017) both present a linear, step-based approach to engineering design. While the literature review indicated that the field has progressed beyond a linear path, Carberry et al (2011) based their study upon state standards at the time. Wendell et al (2017) worked with elementary education students, or beginning learners. As such looking at these items under the lens of beginning learners and past practice, the factor's theme became evident. Each of the items represented the ‘steps’ needed to take a possible solution from an idea to a final, production

solution. With this lens it is clear that factor one represents foundational skills. In addition to being graphical representation rather than a form of modeling, this is the reason that item seven loads onto factor one rather than factor five. Item seven is the creation of production drawings, a skill taught in many introductory graphics and engineering courses. This aligns with the theoretical underpinnings of this study.

The results of the factor analysis were presented in reverse order from lowest variance explained to greatest. During these discussions each factor was also given a name. Each of these factors was discussed with regards to their relationship with the theoretical underpinnings. Each factor was then used to create a composite score for the instrument. As this composite score is meant for interpretation, a discussion of its validity within the engineering design theory is relevant. The composite score was generated by using a weighted average of each factor, with the weights being equal to variance explained. The factor with the most variance explained, foundational skills, was discussed last. Presenting each factor in order of their variance explained sets the stage for the discussion of the composite scores validity. In order they are foundational skills, problem definition, complex problems, experienced engineer skills, advanced knowledge, and pedagogy. As the discussion of the composite score continues it is helpful to remember that engineering design is taught through a series of engineering design problems (Lammi, 2011; Lammi et al., 2018). A series of problems implies the progression of time. As these problems are meant to give students mastery experiences (Lammi, et al., 2018), earlier problems may be shorter and easier allowing for a rapid series of successes to lay a positive foundation of self-efficacy in the learner (Bandura, 1977; Peura, et al, 2021). The largest variation in responses, thus composite scores, was from foundational skills. From a theoretical standpoint this makes sense, teachers who feel they are less able to teach the foundational skills are going to report an

overall lower self-efficacy. These are also the first skills focused on for new learners, as early as elementary education (Wendell et al, 2017). After mastery of teaching foundational skills, the second largest variation is from problem definition. For those that report a high self-efficacy in teaching skills, expanding into teaching more nuanced understandings of the problem definition maps to the first step of the linear engineering design process (Carberry, et al, 2010), an approach also used in elementary education (Wendell, et al, 2017). This remains relevant in the more complex engineering design process found in practice (Lammi et al., 2020) as time spent on problem definition is one of the skills that differentiate a novice and experienced engineering (Wendell, et al, 2017). Understanding how to frame and define the problem opens up the ability to have more complex problems, be they open-ended or multidisciplinary (Lammi, 2011; Lammi et al., 2018). Problem complexity is also the factor that accounts for the third largest variation in responses. As the explained variation decreases, those with lower self-efficacy in teaching engineering design are more likely to have already explained the difference in their scores from one of the previous factors. This leaves only those who have high self-efficacy scores to differentiate between one another. The next factor that explains differences in responses is experienced engineer skills. These all represent skills that differentiate a novice and experienced engineering (Chao, et al, 2017; McFadden and Roehrig, 2019; Wendell, et al., 2017) beyond that of problem definition. This instrument looks at the self-efficacy of teaching these skills. So those with high self-efficacy in this area must believe in their ability to either teach these skills to novice learners or work with advanced learners who perhaps already exhibit a degree of mastery of the previous skills. At this point only two factors remain that differentiate between participants. The remaining two factors are advanced modeling and pedagogy. These two factors each only explain a small portion of the variance. Whereas the previous factors all focused on

increasing complexity of the design problem framework, both of these skills are more general. Modeling techniques receive a large focus starting at the elementary age (Wendell, et. al, 2017) but this continues throughout majority of the engineering design problems (Lammi, et al., 2020; Chao, et al, 2017; Apedoe, et al, 2008; Cunningham, et al, 2020). Pedagogy skills are also more widespread, so those with limited experience teaching engineering design and more experience teaching in other content are likely to have developed these skills previously, leaving only those new to both teaching and teaching engineering design between which to differentiate.

The previous discussion focused on the composite score, which is what this study sought to validate. Given the framing of the discussion, a temptation to look at each factor individually may arise. However, each of these factors have not been validate for individual use. Table 5-1 Factor linear regression shows the results of the linear regression analysis, which indicates not every factor is significantly associated with experience teaching engineering design and experience teaching. The foundational skills and problem identification factors scores do have significant association with experience teaching engineering design and experience teaching. However, only the foundational skills factor was able to keep both as significant predictors at the $p<.05$ level as displayed in Table 5-2 Foundational skills coefficients and Table 5-3 Problem identification coefficients.

Table 5-1.

Factor linear regression

Factor	R	R^2	Adjusted R^2	df	F	Sig
Foundational Skills	0.429	0.184	0.163	2, 78	8.793	<.001
Problem Identification	0.281	0.079	0.055	2, 78	3.341	0.041
Complex problems	0.069	0.005	-0.021	2, 78	0.188	0.829
Experienced engineer skills	0.149	0.022	-0.003	2, 78	0.891	0.414
Advanced modeling	0.033	0.001	-0.025	2, 78	0.043	0.958
Pedagogy	0.218	0.048	0.023	2, 78	1.951	0.149

Table 5-2.

Foundational skills coefficients

Foundational Skills	B	t	Sig
Constant	8.319	27.88	<.001
Eng Des	0.095	4.178	<.001
Gen Teach	-0.054	-2.585	0.012

Table 5-3.

Problem identification coefficients

Problem Identification	B	t	Sig
Constant	8.089	30.359	<.001
Eng Des	0.038	1.852	0.068
Gen Teach	0.001	0.056	0.955

The SMEs indicated that the items have a face validity. The results of the factor analysis are supported by the literature forming the theoretical underpinnings of this study. Both provide

a case for the instruments validity. The linear regression models presented were done so due to the discussion about factors present in this section.

The linear regression models do not indicate that any factor besides foundational skills and problem identification have a significant association with experience teaching engineering design and experience teaching. Moving to a model that only uses experience teaching engineering design retains the significant association of both factors, but also results in experience teaching engineering design ($t=2.601$, $p= .011$) becoming a significant predictor of problem identification score [$F(1,79)=6.765$, $p=.011$, $R^2=.079$)]. However, these results do not weaken the case for validity in regards to engineering design. These results relate to the validity of self-efficacy, thus weaken that case slightly as each factor could not be validated individually.

RQ 4: Instrument validity

The fourth research question was: is there evidence that the instrument, SETED, is valid? The answer to this question lays in the answer to the previous three questions. If the instrument was considered reliable, showed validity within self-efficacy theory, and showed validity within engineering design theory, then it could have been considered valid. Each of the previous three questions were answered in the positive, thus presenting a case for the instrument as a whole to be considered valid. Additionally, there existed a strong, positive correlation between the control item explicitly asking about teaching engineering design and the SETED composite score. The Wilcoxon ranked-signs test revealed that the composite score had a statistically significantly lower mean, although this difference was less than .500. This provides evidence that the instrument follows along with the theoretical underpinnings although it may provide a more nuanced, accurate view of self-efficacy versus a direct question. A one-way ANOVA confirmed no significant effects of gender, which matches the expected theory of experience having an

association. There was an issue with a difference in sample sizes, resulting in Levene's test being performed. The null hypothesis of Levene's test is that the groups compared all have equal population variances. Having failed to find significant results with Levene's test, this study kept the null hypothesis. This strengthens the case of the one-way ANOVA results based upon gender, as the assumption of homogeneity of variance for that analysis likely was not violated.

However, caveats remained. Firstly, only a case for validity can be presented it cannot be guaranteed. The second was that all instrument methodologies showed a second factor analysis process, CFA to confirm the findings of the EFA. The case for validity is not completed until at least a second validating study is completed (Kukul et al, 2019; Sun and Rogers, 2020; Tsai et al, 2021; Xu et al, 2019; Yoon and Evans, 2012; Yoon et al, 2014). While Sun and Rogers (2020) complete this in a single study using a split-half method to break their sample into two smaller groups and running EFA and CFA on separate groups, most studies used a second study towards this end (Tsai et al, 2021; Xu et al, 2019; Yoon and Evans, 2012; Yoon et al, 2014).

Problems

This study faced several issues related to the sample. The first issue was the size of the sample, the goal was to have fifty participants for each group. Only the experienced group came close to this with forty-four participants. As a result, efforts should be made in future research to increase the size of the sample so that each group has at least fifty participants. In addition to the size of the groups, the analyses performed would benefit from having similar or equal group sizes.

In addition to sample size problems, this study intended to have a diverse population of participants. However, due to the recruitment methods majority of the respondents were faculty members at universities. The results may have been skewed as a result of this, because university

faculty tend to have greater experience in teaching or performing engineering design. This study failed to consider the association between self-efficacy in performing engineering design and the self-efficacy of teaching engineering design.

This study did consider the effects of previous teaching experience influencing the self-efficacy of teaching engineering design, but failed to account for an association with performing engineering design. As a result, this demographic data item was missing from the data collection.

The literature review chapter discussed engineering design in terms of STEM education; however, recruitment efforts focused only on the technology and engineering education portions of STEM. The inclusion of science teachers, who may be new to both teaching and performing engineering design, may change the results.

Recruitment for this study occurred during the initial outbreak of COVID-19 in the United States. The disruptions and uncertainty caused by COVID-19 could have led to issues with recruitment efforts as teachers attempted to adapt their curriculums for completely remote learning with no warning. Virtual and electronic forms of communication, such as email, also increased during this time. As recruitment efforts exclusively utilized electronic communications, they may have become lost in the increased volume.

Future research

There are many avenues for which future research can take place. The most basic level is a second, confirmatory study of the results. Future research should focus on increased size and diversity of the sample. In order to increase sample size, the use of recruitment through university graduate schools should be considered. Graduate students within STEM education are more likely to have teaching experience and provide a variety of teaching levels.

The inclusion of graduate students within STEM education would also increase diversity of the sample by including Science education. Future studies should also try and find other methods for including Science educators in the study. Demographic questions already included information for discipline, Science would just need added as an additional option.

Future studies should also consider accounting for the effect of teaching assignment on the results. Educators working at the post-secondary level represent a group of educators with increased experience performing engineering design. This experience of performing engineering design would imply increased self-efficacy in performing engineering design. An increased self-efficacy in performing engineering design may cause an increased self-efficacy in teaching engineering design, skewing the results from those expected. In order to test this adequately, recruitment efforts should focus on obtaining more responses from K-12 teachers. The earlier recommendation of including graduate students may also aid in these efforts.

Demographic data should also include a question about experience in performing engineering design. The inclusion of this question may help to clarify what association, if any, experience performing engineering design has on the results or teaching engineering design.

The previous suggestions all focused on a confirmatory study. However, the instrument also provides the opportunity to expand the population of interest out to Science education. Science education has standards about engineering design, so there should be teachers teaching engineering design within the Science field. A second population, before or after Science education, allows for the inclusion of pre-service teachers. In the current disciplines these pre-service teachers would be within technology education as post-secondary education does not have a pre-service program. It is expected that with the addition of pre-service teachers a wider variety of self-efficacy in teaching engineering design would be obtained. Pre-service teachers

would include students having just entered teacher preparation programs with little to no pedagogical experience as well as a wide variety of experience in teaching engineering design concepts. Expanding the population both ways allows for a much larger pool of participants for a sample, which would provide more conclusive results.

Next steps may also consider first creating and validating an instrument for performing engineering design based upon the theoretical underpinnings presented during this study's literature review. There exist more validated instruments for testing performing engineering design that can be run concurrently with such an instrument. This would align more closely with the guiding study of Carberry et. al. (2010). If such a study were to confirm the validity of the developed instrument and its theoretical underpinnings, this would strength the case for the validity of the instrument presented in this study.

Conclusions

The initial impulse for this study was the researcher's interest in the difference in self-efficacy of teaching engineering design across multiple disciplines within STEM education as well as different levels of experience. The presence of targeted, impactful professional development opportunities for practicing professional relies on the ability to see impacts of the intervention on classroom practices. One method is looking at student achievement scores, although the body of literature also indicates that a teacher's self-efficacy in teaching the subject is also a large determinant of student learning (Cantrell, 2003; Yesilyurt et al, 2016; Cakiroglu et al, 2005). Towards this end this instrument was developed. The goal of this instrument is to identify the differences in self-efficacy of teaching engineering design across all disciplines and across all levels of experience, including pre-service teachers. Having a source of data available to researchers would aid in the development of interventions and professional development

programs for teachers teaching engineering design as finding systematic weakness and strengths would be possible. Since no instrument yet existed, this study sought to begin this process. The complexity of instrument development necessitated that this instrument's scope be reduced to in-service teachers with technology and engineering fields. The efforts of this study were a success, and a case has been made that the SETED is valid within the domains tested. These case for validity would be greatly strengthened by a larger second confirmatory study. There are two possible models that both showed significance. Each model included experience teaching engineering design, but one model also included experience teaching. Both models were close is significance levels, so the single IV model is the most simple and would gain preference. However, due to the nature of engineering design and programs such as Project Lead the Way, which helps to retrain existing teachers in the field of engineering design the second model that utilize total teaching experience is preferred by the researcher because of the wider implications and understanding it provides. A larger percentage of the variance is explained by this model, but it also demonstrates that when combined with engineering design teaching experience total teaching experience has a negative association with the score.

Differences between level of experiences

This finding generates an important discussion through the explanation on the possible causes and implications of a teaching experience mismatch. Those teachers that are, and have always, taught engineering design would have matching years of experience. If each participant had always taught engineering design their entire career the assumption would be that both experience teaching and experience teaching engineering design have the same association with the composite score. Additionally, in a combined model utilizing both forms of teaching experience similar and significant associations would be expected. For total teaching experience

to have no significant association alone, and a negative association when combined with experience teaching engineering design, the implication is that there exist some cases in which a mismatch of experience occurs. A review of the data confirms that there were indeed cases present in which experience teaching and experience teaching engineering design were not the same. These findings ask the questions of why is there a difference between levels of experience and why does this cause negative associations in a combined model? This study postulates, something further occurs with those teachers that have been retrained or added teaching engineering design to their teaching loads later in their careers. This idea is an assumption rooted in the concept that those teachers who have taught engineering design their entire careers would have a one to one match between experience teaching and experience teaching engineering design. Therefore in order to have a mismatch, there exist cases in which a mismatch between experience teaching and experience teaching engineering design would occur. The first case, those that taught engineering design earlier in their career but no longer teach engineering design, were removed from this study by the selection criteria for this study of teachers currently teaching engineering design as a part of their course load. The second case for an experience mismatch was a teacher that previously taught a different subject but has either added or transitioned into teaching engineering design as part of their instructional load. There does exist a third case in which an instructor taught engineering design early in their career, stopped teaching engineering design as part of their course load, and has now returned to teaching engineering design. While a distinct possibility, this study assumes that such cases are minimal and that in any such a case the overall effect on self-efficacy of teaching engineering design would also be minimal. Overall the trend still matches the theory with the composite score trending positively with years of experience, allowing the instrument to remain valid. The

intersection between experiences that results in cases of mismatched experience is the reason this study favors the use of a model that considers both types of teaching experience moving forward.

Straight average regression model

The previous discussion was generated from a comparison of possible models. In reviewing the methodology and results, a question raises as to why weighted scores were used over straight averages. While this study presented a case for the use of weighted composite scores (Glen, 2015; Starkweather, 2012; Malandrakis et al., 2019; Tasi et al, 2018; Tsai et al, 2021; Xu et al, 2019) in the interest of a transparent discussion the same regression tests using experience teaching engineering design, teaching experience, and both variables were completed. The results are found in Table 5-4 Straight average regression models.

Table 5-4.

Straight average regression models

Model	R	R^2	Adjusted R^2	df	F	Sig
1	0.208	0.043	0.031	1, 79	3.581	0.062
2	0.044	0.002	-0.011	1, 79	0.151	0.699
3	0.248	0.061	0.037	2, 78	2.555	0.084

These results show that using a straight average that no model shows a significant association with experience teaching engineering design, experience teaching, or both.

Use of a single item

The results of this study are not without their own merits. This study used a control item asking simply in the participants belief in their ability to teach students engineering design. Linear regression using the same two experience model score [$F(2,78)=5.495$, $p=.006$, $R^2=.124$)] as the composite score showed significant association between ED response scores and experience teaching engineering design ($t=3.300$, $p=.001$) and experience teaching ($t=-2.020$,

p=.047), remaining consistent with the self-efficacy underpinnings. Such a question is used on instruments measuring the self-efficacy in teaching engineering (Yoon et al, 2014). The theoretical underpinnings of this study postulate that engineering design is complex (Lammi, 2011) and thus a single item is an over-simplification of the engineering design concept. The Wilcoxon signed-rank test revealed that the control item had a larger range in responses when compared to the composite score. Both scores had a maximum of ten, but the control items range was lower with a minimum of 2 compared to the composite score minimum of 5.46. Even given the wider range with a lower minimum bound of the control item, the Wilcoxon test revealed that there were more negative cases, or instances where the composite score was lower than the reported score on the control item. This information is of limited use alone, as the control item was restricted to whole numbers due to the ordinal scale whereas the composite score functioned on a continuous scale. If the assumption that the composite score is indeed valid holds true, when looking at the Wilcoxon results a few pieces of information become evident. The first revelation is that given the higher minimum of the composite score and 24 positive rank cases, those with lower self-efficacy are more likely to underestimate their ability and report a lower self-efficacy when asked directly their self-efficacy of teaching engineering design. By the same token those with moderate to high self-efficacy are more likely to overestimate their abilities.

Factor analysis

The factor analysis revealed six underlying factors. Each of these factors could be grouped thematically and explained as a cohesive unit. The variances explained by each factor provide a road map for instructional practices. It is clear that the largest differentiation between a teacher with high and moderate self-efficacy is their beliefs in their ability to teach the foundational skills of engineering design. The next largest group relates to the teacher's ability to

create and provide scaffolding during complex problems, a theme evident from the literature review as well. There is a clear progression into the more advanced teaching engineering design skills. The least effect on the self-efficacy of teaching engineering design was the teacher's ability to create assessment tools. This is likely a result of only a single item related to pedagogy and assessment. A more through literature review of assessing engineering design and providing scaffolded instruction to create more pedagogy items would be of benefit, but the instrument then begins to drift into general teaching pedagogy rather than domain specific pedagogy. The results also indicated mean values above the mid-way point indicating that all the teachers surveyed had moderate to high self-efficacy.

AERA Standards

In a previous chapter twenty-four standards were selected from the AERA (2014) *Standards for Educational and Psychological Testing*. These standards represent a small cross section of the total available standards. While the *Standards* address the fact that not every standard will apply to every instrument, the remained a significant number of relevant standards that were unaddressed by this study. As a result, further research and work is needed before the instrument could be considered completed. However, this study did address the previously identified standards. Reviewing each of these and how they were addressed by this study helps to demonstrate the existing foundation upon which work can continue.

This study presented a clear methodological base for the generation of items, reviewed by subject matter experts, upon which the instrument was formed. The selection criteria and process for the subject matter experts could have been better delineated, but the process used was adequate enough that the results of the expert feedback remained valid. The instrument was presented via technology and use of the internet, but the items on the instrument were not of the

type in which older technology would be presented with a disadvantage. As time was not a factor considered in the analysis of the instrument, any delay due to a slower connection or device was also not a concern for this study. The instrument was presented in English, the dominant language for the intended population. The methodology had a clear process for analyses to find and report evidence for validity and reliability. A discussion of this evidence occurs earlier in this chapter. Several problems, also discussed previously, do weaken the evidence presented, but not to the point where the achievement of a solid case for validity is unachievable by future research. The major failing thus far of this study is in the interpretation of the results. As the instrument is not validated, an interpretation of the results would be premature. However, the nature of the test itself does highlight a few key points.

The composite score generated by the SETED should not be treated as an absolute value, where the single number on the 1-10 scale has meaning. Rather the score would be used to compare individuals to one another or an individual to a group of peers. An individual's score could also be compared to their own score in a pre-post format involving an intervention. Depending on the testing interval, experience teaching engineering design could be presented as a controlled variable. As this study found that experience teaching engineering design had the strongest association with an increase in composite scores, such a format would allow for any significant change to be attributed to the intervention itself. The impetus of this study was to develop an instrument to evaluate an individual's self-efficacy of teaching engineering design. As a part of this intent, the goal was for the resulting instrument to simply inform where an individual is. Holistically this thought is that of 'where we are' and the ability to use the knowledge of where we presently are to inform next steps. While there are theoretical relationships between teaching self-efficacy and effective teachers, the score has not been

validated for this use, is not intended for such use, and should not be treated as a method for evaluation of the quality of a teacher. Each item was not individually validated, nor were each of the factors, as a result the instrument should only be used to report a single final score.

Composite score meaning

Thus far no discussion has been given to the meaning of the composite score. Self-efficacy has several positive correlations with desirable attributes in teachers and their pupils. Those with higher self-efficacy are more likely to persist in the face of setbacks (Tschannen-Moran and Hoy, 2001), are more willing to implement new instructional practices (Dussault, 2006), and discuss short comings with others (Lammi, et al, 2018). More experienced teachers and those with higher self-efficacy are more prepared to offer the necessary scaffolding (Chao, et al, 2017), help students cope with failure (Aydeniz, et al, 2020; Lammi, et al, 2018; Wendell, et al, 2017) and improve pupil's attitude towards content (Cantrell, 2003; Yesilyurt et al, 2016). Students learn more from those teachers with higher self-efficacy (Cakiroglu, et al, 2005) and are more likely to achieve higher scores on standardized tests (Cantrell, 2003). This score is not an guarantee of these attributes, it simply indicates that they are more likely to occur. Just as the composite score itself may vary, the effects of self-efficacy between people may also vary. The likelihood of these attributes being positively displayed are more likely as the score increases, but that does not mean those at the lower end of scores will not display any of these attributes.

Earlier in the discussion it was noted that a Wilcoxon ranked sign test revealed several instances that when directly asked participants reported a self-efficacy lower than that calculated by this instrument. Based upon this disconnect the expectation when using a singular item would be that those reporting a lower self-efficacy on this item would be less likely to display these positive attributes. However, the composite score reveals that when a nuanced look at teaching

engineering design is taken, those reporting a lower self-efficacy on a single item may be more likely to display these positive attributes based upon their composite score.

Summary

A next study is required to further increase the case for validity. Such a study should focus on a larger, more diverse sample that better aligns with the intended population. A confirmatory analysis of the factor structure is needed along with further validation of the regression model presented. The instrument as it currently exists should not be used in the field, as further pilot testing is required. A caution about use of any such validated instrument is issued. While there exist several different factors, these factors have not been validated for individual use so all items should be used and a composite score generated. This instrument is not meant for evaluative purpose, only informational purposes. The most value to be gained from the reported score is through relative comparisons. These comparisons may be between teachers, but should not be evaluative in nature. The intent is that this scale is to be used across time intervals with the same teacher, such as before and after an intervention. Given the correlations between self-efficacy of teachers, those teachers positive habits, and students knowing a teachers self-efficacy of teaching engineering design helps to understand where additional issues may be expected. This study makes a case for use of the generated composite score makes a case for giving a more complete picture of self-efficacy of teaching engineering design as opposed to a single item asking a respondent to report their own self-efficacy of teaching engineering design.

REFERENCES

ABET. (2012a). *2013-2014 Criteria for accrediting engineering programs.* Baltimore, MD

ABET. (2012b). *2013-2014 Criteria for accrediting engineering technology programs.* Baltimore, MD

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

American Society of Engineering Educators. (2013). Our Vision. Retrieved from http://www.asee.org/about-us/the-organization/our-mission.

Apedoe, X., Reynolds, B., Ellefson, M., and Schunn, C. (2008). Bringing engineering design into high school science classrooms: The heating/cooling unit. *Journal of Science Education and Technology, 17*(5), 454-465.

Asunda, P. A., and Hill, R. B. (2007). Critical features of engineering design in technology education. *Journal of Industrial Teacher Education, 44*(1), 25-48.

Aydeniz, M., Bilican, K., Senler, B. (2021) Development of the inquiry-based science teaching efficacy scale for primary teachers. *Science & Education, 30,* 103-120.

Baker, D., Krause, S., and Purzer, S. (2008). *Developing an instrument to measure tinkering and technical self-efficacy in engineering.* Paper presented at the 2008 Annual Conference of American Society of Engineering Educators, Pittsburg, PA.

Bandura, A. (1977). Self-efficacy: towards a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologists*, *37*(2), 122-47.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents.* (307-337)

Board of Science Education. (2011). A framework for K-12 Science Education: practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press.

Boser, U. (2012) Race to the top: What have we learned from the states so far? Washington, DC: Center for American Progress

Brophy, S., Klein, S., Portsmore, M., and Rogers, C. (2008). Advancing engineering education in P-12 classrooms. *Journal of Engineering Education, 97*(3), 369-387.

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *Japan Association for Language Teaching Testing & Evaluation SIG Newsletter, 13 (3), 20-25.*

Çakiroglu, J., Çakiroglu, E., and Boone, W. J. (2005). Pre-service teacher self-efficacy beliefs regarding science teaching: a comparison of pre-service teachers in Turkey and the USA. *Science Educator*, 14(10), 31-40.

Cantrell, P. (2003). Traditional vs. retrospective pretests for measuring science teaching efficacy beliefs in preservice teachers. *School Science and Mathematics,* 103(4), 177-185.

Carberry A. R., Lee HS., and Ohland M. (2010). Measuring engineering design self-efficacy. *Journal of Engineering Education*, 99(1),71-79.

Carr, R. L., and Strobel, J. (2011). Integrating engineering design challenges into secondary STEM education. Logan, UT: National Center for Engineering and Technology Education, Utah State University.

Chao, J., Xie, C., Nourian, S., Chen, G., Bailey, S., Goldstein, M. H., Purzer, S., Adams, R. S., Tutwiler, M. S. (2017) Bridging the design-science gap with tools: science learning and design behaviors in a simulated environment for engineering design. *Journal of Research in Science Teaching*, 54(8), 1049-1096.

Clark, A. C. (2012). "Excellence" in STEM education. *Technology & Engineering Teacher,* 72(2), 33-36.

Cohen, Jacob (1977). *Statistical power analysis for the behavioral sciences.* Academic Press, Inc.

Concannon, J. P. and Barrow, L. H. (2009). A cross-sectional study of engineering students' self-efficacy by gender ethnicity, year, and transfer status. *Journal of Science Education Technology,* 18, 163-172.

Cunningham, C. M., Lachapelle, C. P., Brennan, R. T., Kelly, J. K., Tunis, C. S. A., Gentry, C. A., (2020) The impact of engineering curriculum design principles on elementary students' engineering and science learning. *Journal of Research in Science Teaching*, 75, 423-453

Dellinger A. B., Bobbett J. J., Olivier D.F., and Ellet, C.D (2008). Measuring teachers' self-efficacy beliefs: Development and use of the TEBS-Self. *Teaching and Teacher Education,* 24, 751-766.

Dussault, M. (2006). Teachers' self-efficacy and organizational citizenship behaviors, *Psychological reports*, 98, 427-432.

Elliot, A. J., Murayama, K., and Pekrun, R. (2011). A 3 X 2 achievement goal model. *Journal of Educational Psychology*. 103(3), 632-648.

Erdem, E., Demirel, O. (2007). Teacher self-efficacy belief. *Social Behavior and Personality*, 35(5), 573-586.

Eisenkraft, A. (2011). Engineering design challenges in a science curriculum. Logan, UT: National Center for Engineering and Technology Education, Utah State University.

Fantz, T.D., Siller, T. J., and DeMiranda, M. A. (2011). Pre-collegiate factors influencing the self-efficacy of engineering students. *Journal of Engineering Education,* 100(3), 604-623.

Falco, L. D. and Summers, J. J., (2019). Improving career decision self-efficacy and STEM self-efficacy in high school girls: evaluation of an intervention. *Journal of Career Development,* 46(1), 62-76.

Finnegan, R. S. (2013). Linking teacher self-efficacy to teacher evaluations. *Journal of Cross-Disciplinary Perspectives in Education,* 6(1), 18-25.

Gist, M. E., Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *Academy of Management*, 17(2), 183-211.

Glen, S. (2015). Likert scale definition and examples. *Statisticshowto.com: elementary statistics for the rest of us!*

Gray, L., Taie, S. (2015). Public school teacher attrition and mobility in the first five years: results from the first through fifth waves of the 2007-08 beginning teacher longitudinal study. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

International Technology and Engineering Educators Association. (2007). *Standards for technological literacy* (3rd ed.). Reston, VA.

International Technology and Engineering Educators Association. (2020). *Standards for technological and engineering literacy.* Reston, VA.

Jimoyiannis, Athanassios and Komis, Vassilis. (2007). *Examining teachers' beliefs about ICT in education: Implications of a teacher preparation programme.* Teacher Development. 11. 149-173. 10.1080/13664530701414779.

Jonassen, D. (2011). Design problems for secondary students. Logan, UT: National Center for Engineering and Technology Education, Utah State University.

Kukul, V., Karatas, S. (2019) Computational thinking self-efficacy scale: Development, Validity, and Reliability. *Informatics in Education.* 18(1), 151-164.

Laerd Statistics. (2018). Kruskal-Wallis H Test using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

Laerd Statistics. (2018). Linear regression Analysis using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

Laerd Statistics. (2018). Mann-Whitney U Test using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

Laerd Statistics. (2018). One-way ANOVA in SPSS Statistics. https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Laerd Statistics. (2018). Pearson's Product-Moment Correlation using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php

Laerd Statistics. (2018). Testing for Normality using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

Lammi, M. D. (2011). *Engineering Design Challenges*. Unpublished manuscript, Department of Science, Technology, Engineering, and Mathematics Education, North Carolina State University, Raleigh, N.C.

Lammi, M. D., Wells, J. G., Gero, J. (2020). High school pre-engineering students' engineering design cognition. *International Journal of Technology and Design Education*

McFadden, J., Roehrig, G. (2019). Engineering design in the elementary science classroom: supporting student discourse during an engineering design challenge. *International Journal of Technology and Design Education.,* 29, 231-262

National Assessment of Educational Progress. (2014). *2014 Abridged Technology and Engineering Literacy Framework*. Washington, DC: The National Academies Press.

National Research Council. (2013). *Next generation science standards.* Washington, DC: The National Academies Press.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational research,* 66(4), 543-578.

Panadero, E., Jonsson, A., Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: four meta-analyses. *Educational research review,* 22, 74-98.

Peura, P., Tuija A., Raikkonen, E., Viholainen, H., Koponen, T., Usher, E. L., Aro M. (2021). Trajectories of change in reading self-efficacy: a longitudinal analysis of self-efficacy and its sources. *Contemporary educational psychology*., 64.

Pouta, M., Lehtinen, E., Palonen, T. (2020). Student Teachers' and Experienced Teachers' professional vision of students' understanding of the rational number concept. *Educational Psychology Review*.,

Ritter, J. M., Boone, W. J., and Rubba, P. A. (2001). Development of an instrument to assess prospective elementary teacher self-efficacy beliefs about equitable science teaching and learning (SEBEST). *Journal of Science Teacher Education,* 12(3), 175-198.

Sadler, P. M., Coyle, H. P., and Schwartz, M. (2000). Engineering competitions in the middle school classroom: Key elements in developing effective design challenges. *Journal of the Learning Sciences, 9*(3), 299 - 327.

Silk, E. M., and Schunn, C. (2008). *Core concepts in engineering as a basis for understanding and improving K-12 engineering education in the United States*. Paper presented at the National Academy Workshop on K-12 Engineering Education, Washington, DC.

Smolleck, L. D., Zembal-Saul, C., and Yoder, E. P. (2006). The development and validation of an instrument to measure preservice teachers' self-efficacy in regard to the teaching of science as inquiry. *Journal of Science Teacher Education*, 17, 137-163.

Starkweather, J. (2012). How to calculate empirically derived composite or indicate scores. *Journal of data science,* February.

Sun, Y., Rogers, R. (2020) Development and validation of the online learning self-efficacy scale (OLSS): A structural equation modeling approach. *American Journal of Distance Education*

Tavakol, M., and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education,* 2, 53-55.

Tsai, M. J., Wang, C. Y., Hsu, P. F. (2019) Developing the computer programming self-efficacy scale for computer literacy education. *Journal of educational computing research,* 56(8), 1345-1360.

Tsai, M. J., Wang, C. Y., Wu, A. H., Hsiao, C. Y. (2021) The development and validation of the robotics learning self-efficacy scale. *Journal of educational computing research,* 0(0), 1-19.

Tschannen-Moran, M., Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17, 783-805.

Tschannen-Moran, M., Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education,* 23, 944-956.

UCLA: Statistical Consulting Group. (n.d.) *Factor analysis | SPSS annotated output*. https://stats.idre.ucla.edu/spss/output/factor-analysis/

UCLA: Statistical Consulting Group. (n.d.). *A practical introduction to factor analysis: exploratory factor analysis.* https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/

Wendell, K. B., Wright, C. G., Paugh, P. (2017). Reflective decision-making in elementary students' engineering design. *Journal of Engineering Education*, 106(3), 356-397.

Xu, M. Williams, J. P., Gu, J. (2020) An initial development and validation of a Chinese technology teachers' attitudes towards technology (TTATT) scale. *International Journal of Technology and Design Education*. 3-, 937-950

Yesilyurt, E., Ulas, A. H., Akan, D. (2016) Teacher self-efficacy, academic self-efficacy, and computer self-efficacy as predictors of attitude toward applying computer-supported education. *Computers in human behavior,* 64, 591-601.

Yoon, S. Y., Evans, M. G. (2012). *Development of the teaching engineering self-efficacy scale (TESS) for k-12 teachers*. 2012 Annual Conference of American Society of Engineering Educators, San Antonio, TX.

Yoon, S. Y., Evans, M. G., Strobel, J. (2014). Validation of the teaching engineering self-efficacy scale for K-12 teachers: a structural equation modeling approach. *The research journal for engineering education,* 103(3), 463-485.

Zee, M., Koomen, H. M. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: a synthesis of 40 years of research. *Review of educational research*, 86(4), 981-1015.

APPENDICES

Appendix A: Items

Item number	Item wording
1	I feel I can successfully teach students to engage in open-ended problems (e.g. no known solution, multiple solutions, multiple paths to a solution).
2	I am comfortable helping students work a multidisciplinary problem.
3	I feel I can help students successfully engage in systems thinking.
4	I believe I can teach students how to produce a working model. (e.g. proof of concept, prototype, mock-up)
5	I believe I can teach students how to use simulations to represent a design and the associated concepts.
~~6~~	~~I feel I can teach students how to visually render their ideas. (e.g. hand-sketches, drawings, 3D computer models)~~
7	I believe I can teach students how to work with production drawings and spec sheets.
8	I feel I can teach students how to make performance based trade-offs.
9	I feel I can teach students how to design under constraint.
10	I believe I can teach students to identify when a solution satisfies the problem.
11	I feel I can teach students how to formulate a problem.
12	I believe I can teach students to optimize a design.
13	I feel I can teach students how to ask purposeful questions throughout the design process.
14	I believe I can teach students how to create a working definition of the problem.
15	I believe I can teach students to analyze, evaluate, and make changes to their solutions as they work.
16	I feel I can teach students to apply science principles and practices to a problem.
17	I feel I can teach students to think mathematically to understand and solve a problem.
~~18~~	~~I believe I can teach students how to work with limited resources.~~
~~19~~	~~I believe I can teach students to appropriately use, modify, and test available materials and resources to produce a physical representation of an idea.~~
20	I feel I can teach students to make cost-based trade offs.
21	I feel that I can help students deal with ambiguous constraints.
22	I believe I can create and use metrics to assess students' design and engineering work successfully.
~~23~~	~~I feel I can teach problem solving that is relevant and significant to students' lives.~~
24	I feel I can help students work real-world engineering problems.
25	I feel I can teach students to be productive members of a design team.
26	I believe I can teach students that "failure" is a part of the design process.

27	I feel I can teach students that design is a continual, cyclical process.
28	I feel I can teach students to effectively communicate ideas in a written or verbal form.
29	I believe I can teach students engineering design.

Appendix B: Pearson's Correlation

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.00												
2	.497**	1.00											
3	.590**	.670**	1.00										
4	.566**	.351**	.482**	1.00									
5	.247*	.103	.275*	.462**	1.00								
6	.230*	.002	.318**	.374**	.484**	1.00							
7	.293**	-.032	.301**	.307	.276*	.609**	1.00						
8	.401**	.145	.467**	.369**	.272*	.469**	.647**	1.00					
9	.494**	.149	.501**	.436**	.207	.437**	.640**	.818**	1.00				
10	.563**	.302**	.459**	.423**	.108	.144	.321**	.610**	.620**	1.00			
11	.345**	.489**	.568**	.308**	.137	.293**	.257*	.473**	.356**	.457**	1.00		
12	.454**	.257*	.489**	.375**	.168	.487**	.547**	.668**	.682**	.502**	.670**	1.00	
13	.392**	.265*	.394**	.318**	.192	.273*	.347**	.574**	.458**	.478**	.493**	.456**	1.00
14	.296**	.255*	.397**	.294**	.270*	.363**	.471**	.629**	.401**	.473**	.602**	.455**	.750**
15	.383**	.283*	.437**	.330**	.165	.149	.244*	.469**	.385**	.564**	.515**	.517**	.584**

*p<.05. **p<.01. ***p<.001

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
16	.360**	.439**	.439**	.207	.066	-.009	.185	.491**	.388**	.638**	.469**	.456**	.586**
17	.390**	.476**	.476**	.387**	.226*	.254*	.273*	.326**	.389**	.481**	.454**	.568**	.430**
18	.365**	.165	.265*	.370**	.098	.303**	.358**	.475**	.513**	.547**	.369**	.513**	.499**
19	.260*	.132	.284*	.378**	.099	.225*	.424**	.362**	.341**	.309**	.400**	.436**	.418**
20	.336**	.201	.519**	.289**	.139	.394**	.503**	.793**	.641**	.460**	.600**	.731**	.539**
21	.391**	.281*	.527**	.343**	.221*	.469**	.513**	.634**	.593**	.524**	.671**	.771**	.639**
22	.238*	.225*	.558**	.333**	.323**	.422**	.313**	.358**	.318**	.177	.495**	.379**	.317**
23	.198	.128	.394**	.204	.186	.204	.235*	.327**	.355**	.314**	.298**	.366**	.498**
24	.482**	.221*	.483**	.364**	.126	.430**	.638**	.722**	.776**	.544**	.385**	.686**	.489**
25	.335**	.345**	.337**	.375**	.370**	.425**	.198	.192	.193	.155	.364**	.302**	.332**
26	.274*	.291**	.207	.111	.112	.257*	.331**	.301**	.322**	.202	.258*	.300**	.408**
27	.295**	.164	.316**	.309**	.166	.332**	.378**	.485**	.512**	.322**	.286**	.445**	.549**
28	.371**	.328**	.458**	.496**	.291**	.403**	.418**	.380**	.334**	.194	.374**	.365**	.483**
29	.404**	.121	.299**	.404**	.229*	.548**	.583**	.512**	.599**	.345**	.362**	.587**	.396**

*p<.05. **p<.01. ***p<.001

Variable	14	15	16	17	18	19	20	21	22	23	24	25	26
14	1.00												
15	.655**	1.00											
16	.576**	.634**	1.00										
17	.349**	.473**	.656**	1.00									
18	.522**	.479**	.404**	.254*	1.00								
19	.398**	.464**	.289**	.260*	.498**	1.00							
20	.583**	.517**	.491**	.426**	.491**	.466**	1.00						
21	.652**	.596**	.501**	.542**	.555**	.568**	.716**	1.00					
22	.337**	.320**	.311	.435**	.162	.302**	.486**	.485**	1.00				
23	.408**	.484**	.467**	.310**	.500**	.379**	.478**	.445**	.512**	1.00			
24	.399**	.317**	.404**	.451**	.447**	.323**	.719**	.619**	.399**	.368**	1.00		
25	.330**	.272*	.173	.289**	.392**	.368**	.292**	.56**	.421**	.328**	.224*	1.00	
26	.328**	.264*	.142	.135	.443**	.294**	.353**	.247*	.106	.399**	.319**	.556**	1.00
27	.438**	.370**	.332**	.248*	.629**	.439**	.518**	.492**	.306**	.593**	.478**	.481**	.626**
28	.528**	.334**	.272*	.338**	.506**	.379**	.452**	.483**	.453**	.392**	.330**	.663**	.433**
29	.380**	.172	.175	.293**	.454**	.390**	.427**	.483**	.228*	.214	.694**	.359**	.367**

*p<.05. **p<.01. ***p<.001

Variable	27	28	29
27	1.00		
28	.655**	1.00	
29	.576**	.634**	1.00

*p<.05. **p<.01. ***p<.001

Appendix C: Spearman's Correlations

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.000												
2	.542**	1.000											
3	.564**	.694**	1.000										
4	.626**	.465**	.531**	1.000									
5	.412**	.255*	.355**	.559**	1.000								
6	.307**	.096	.361**	.491**	.529**	1.000							
7	.378**	.131	.375**	.466**	.396**	.589**	1.000						
8	.426**	.342**	.549**	.460**	.398**	.443**	.606**	1.000					
9	.463**	.267*	.527**	.500**	.325**	.390**	.634**	.825**	1.000				
10	.552**	.485**	.512**	.539**	.258*	.243*	.444**	.606**	.638**	1.000			
11	.450**	.592**	.614**	.406**	.227*	.326**	.350**	.554**	.448**	.555**	1.000		
12	.462**	.408**	.496**	.459**	.303**	.432**	.517**	.680**	.631**	.542**	.718**	1.000	
13	.559**	.436**	.491**	.442**	.279*	.281*	.396**	.615**	.512**	.601**	.627**	.570**	1.000
14	.491**	.430**	.489**	.354**	.293**	.379**	.479**	.624**	.519**	.535**	.672**	.556**	.783**
15	.468**	.487**	.539**	.418**	.263*	.210	.347**	.504**	.464**	.645**	.550**	.529**	.667**

*p<.05. **p<.01. ***p<.001

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
16	.405**	.554**	.598**	.330**	.213	.062	.257*	.573**	.486**	.687**	.583**	.578**	.623**
17	.352**	.463**	.566**	.437**	.359**	.342**	.402**	.458**	.44**	.538**	.514**	.587**	.455**
18	.459**	.342**	.296**	.474**	.179	.243*	.397**	.433**	.495**	.583**	.407**	.502**	.534**
19	.375**	.303**	.389**	.513**	.191	.276*	.491**	.453**	.419**	.425**	.466**	.433**	.536**
20	.379**	.96**	.581**	.421**	.238*	.302**	.473**	.803**	.705**	.547**	.563**	.722**	.645**
21	.379**	.387**	.573**	.438**	.294**	.425**	.583**	.660**	.586**	.500**	.683**	.688**	.663**
22	.331**	.333***	.537**	.420**	.383**	.411**	.394**	.429**	.366**	.300**	.487**	.443**	.376**
23	.300**	.283*	.427**	.340**	.290**	.228*	.320**	.410**	.454**	.415**	.350**	.465**	.558**
24	.454**	.341**	.506**	.499**	.268*	.353**	.602**	.679**	.698**	.587**	.438**	.617**	.576**
25	.380**	.389**	.424**	.394**	.356**	.332**	.211	.270*	.264*	.277*	.412**	.413**	.403**
26	.342**	.312**	.319**	.254*	.199	.209	.416**	.360**	.431**	.374**	.308**	.371**	.461**
27	.378**	.320**	.436**	.444**	.217	.262*	.400**	.522**	.565**	.496**	.388**	.496**	.596**
28	.438**	.404**	.534**	.544**	.321**	.384**	.409**	.435**	.390**	.267*	.488**	.489**	.514**
29	.535**	.279**	.464**	.548**	.418**	.558**	.529**	.451**	.532**	.470**	.563**	.576**	.501**

*p<.05. **p<.01. ***p<.001

Variable	14	15	16	17	18	19	20	21	22	23	24	25	26
14	1.000												
15	.654**	1.000											
16	.628**	.717**	1.000										
17	.419**	.511**	.676**	1.000									
18	.557**	.504**	.520**	.355**	1.000								
19	.490**	.539**	.401**	.348**	.478**	1.000							
20	.617**	.622**	.611**	.523**	.486**	.529**	1.000						
21	.632**	.572**	.555**	.527**	.524**	.618**	.754**	1.000					
22	.364**	.406**	.367**	.522**	.292**	.419**	.453**	.570**	1.000				
23	.537**	.609**	.602**	.429**	.543**	.411**	.584**	.464**	.524**	1.000			
24	.485**	.508**	.522**	.516**	.515**	.461**	.718**	.653**	.469**	.553**	1.000		
25	.412**	.369**	.331**	.360**	.393**	.428**	.383**	.438**	.513**	.424**	.355**	1.000	
26	.424**	.462**	.296**	.319**	.423**	.330**	.474**	.334**	.247*	.507**	.407**	.529**	1.000
27	.481**	.492**	.497**	.404**	.563**	.491**	.600**	.522**	.408**	.643**	.569**	.507**	.652**
28	.554**	.399**	.397**	.398**	.430**	.420**	.506**	.581**	.513**	.433**	.405**	.651**	.420**
29	.475**	.353**	.347**	.468**	.466**	.436**	.397**	.519**	.412**	.393**	.591**	.463**	.393**

*p<.05. **p<.01. ***p<.001

Variable	27	28	29
27	1.000		
28	.466**	1.000	
29	.433**	.418**	1.000

*p<.05. **p<.01. ***p<.001

Appendix D: Factor pattern matrix

	Factor Loading					
Item	1	2	3	4	5	6
1	0.243	0.068	**0.554**	0.091	0.218	-0.244
2	-0.219	-0.077	**0.908**	0.185	-0.084	0.090
3	0.175	-0.028	**0.729**	-0.045	0.075	0.277
4	0.050	0.068	0.302	-0.044	**0.621**	-0.133
5	-0.088	0.079	-0.007	-0.130	**0.696**	0.126
6	**0.450**	-0.180	-0.188	-0.004	**0.552**	0.297
7	**0.750**	-0.067	-0.257	0.020	0.252	0.066
8	**0.775**	0.243	-0.127	-0.076	0.066	-0.026
9	**0.965**	-0.123	0.061	0.056	0.006	-0.157
10	0.345	**0.498**	0.223	-0.116	-0.007	-0.369
11	0.059	**0.412**	0.283	-0.040	-0.027	0.337
12	**0.655**	0.111	0.129	-0.005	-0.054	0.186
13	-0.003	**0.716**	-0.059	0.189	0.027	-0.024
14	-0.088	**0.911**	-0.213	0.003	0.195	0.067
15	-0.163	**0.916**	0.022	-0.035	0.039	-0.025
16	-0.056	**0.769**	0.320	-0.108	-0.221	0.000
17	0.094	0.256	**0.479**	-0.155	0.045	0.216
18	0.166	**0.438**	-0.117	**0.464**	-0.005	-0.260
19	0.078	0.386	-0.09	0.227	0.074	0.040
20	**0.619**	0.218	0.001	0.105	-0.183	0.268
21	0.342	**0.491**	-0.008	-0.025	0.048	0.255
22	0.074	0.076	0.207	-0.039	0.202	**0.572**
23	-0.010	0.397	-0.011	0.372	-0.111	0.140
24	**0.964**	-0.223	0.159	0.098	-0.139	0.035
25	-0.272	-0.038	0.236	**0.612**	0.303	0.193
26	0.118	-0.140	0.090	**0.865**	-0.183	-0.089
27	0.229	0.110	-0.059	**0.740**	-0.118	-0.069
28	-0.112	0.143	0.121	**0.456**	0.314	0.157

Note. N=81. The extraction method was principal axis factoring with an oblique (Promax with Kaiser Normalization) rotation. Factor loadings above .400 are in bold.

Appendix E: Items by factor

Factor 1: Foundational skills

Item	Item wording	Loading
9	I feel I can teach students how to design under constraint.	.965
24	I feel I can help students work real-world engineering problems.	.964
8	I feel I can teach students how to make performance based trade-offs.	.775
7	I believe I can teach students how to work with production drawings and spec sheets.	.750
12	I believe I can teach students to optimize a design.	.655
20	I feel I can teach students to make cost-based trade offs.	.619
~~6~~	~~I feel I can teach students how to visually render their ideas. (e.g. hand-sketches, drawings, 3D computer models)~~	.450

Factor 2: Problem identification

Item	Item wording	Loading
15	I believe I can teach students to analyze, evaluate, and make changes to their solutions as they work.	.916
14	I believe I can teach students how to create a working definition of the problem.	.911
16	I feel I can teach students to apply science principles and practices to a problem.	.769
13	I feel I can teach students how to ask purposeful questions throughout the design process.	.716
10	I believe I can teach students to identify when a solution satisfies the problem.	.498
21	I feel that I can help students deal with ambiguous constraints.	.491
~~18~~	~~I believe I can teach students how to work with limited resources.~~	~~.438~~
11	I feel I can teach students how to formulate a problem.	.412

Factor 3: Complex engineering design problem

Item	Item wording	Loading
2	I am comfortable helping students work a multidisciplinary problem.	.908
3	I feel I can help students successfully engage in systems thinking.	.729
1	I feel I can successfully teach students to engage in open-ended problems (e.g. no known solution, multiple solutions, multiple paths to a solution).	.554
17	I feel I can teach students to think mathematically to understand and solve a problem.	.479

Factor 4: Experienced engineer skills

Item	Item wording	Loading
26	I believe I can teach students that "failure" is a part of the design process.	.865
27	I feel I can teach students that design is a continual, cyclical process.	.740
25	I feel I can teach students to be productive members of a design team.	.612
~~18~~	~~I believe I can teach students how to work with limited resources.~~	~~.464~~
28	I feel I can teach students to effectively communicate ideas in a written or verbal form.	.456

Factor 5: Advanced modeling

Item	Item wording	Loading
5	I believe I can teach students how to use simulations to represent a design and the associated concepts.	.696
4	I believe I can teach students how to produce a working model. (e.g. proof of concept, prototype, mock-up)	.621
~~6~~	~~I feel I can teach students how to visually render their ideas. (e.g. hand-sketches, drawings, 3D computer models)~~	.552

Factor 6: Pedagogy

Item	Item wording	Loading
22	I believe I can create and use metrics to assess students' design and engineering work successfully.	.572

Appendix F: Cronbach's α

Item	Cronbach's α with removal
Base	0.947
1	0.945
2	0.947
3	0.944
4	0.946
5	0.948
~~6~~	~~0.947~~
7	0.945
8	0.943
9	0.944
10	0.945
11	0.945
12	0.943
13	0.944
14	0.944
15	0.945
16	0.945
17	0.945
~~18~~	~~0.950~~
~~19~~	~~0.946~~
20	0.943
21	0.943
22	0.946
~~23~~	~~0.946~~
24	0.944
25	0.946
26	0.946
27	0.945
28	0.945
29	0.945

Appendix G: Factor Correlations

Factor	ED1	ED2	ED3	ED4	ED5	ED6	ED	EDC
ED1	1.000							
ED2	.679**	1.000						
ED3	.477**	.610**	1.000					
ED4	.484**	.501**	.432**	1.000				
ED5	.359**	.327**	.421**	.365**	1.000			
ED6	.432**	.430**	.441**	.370**	.380**	1.000		
ED	.666**	.398**	.320**	.483**	.347**	.228*	1.000	
EDC	.968**	.779**	.629**	.596**	.481**	.560**	.655**	1.000

*p<.05. **p<.01. ***p<.001

Appendix H: Linear regression results

Model	R	R^2	Adjusted R^2	df	F	Sig
1	0.307	0.094	0.083	1,79	8.232	0.005
2	0.049	0.002	-0.01	1,79	0.194	0.661
3	0.377	0.142	0.12	2,78	6.469	0.003

Model 1 uses experience teaching engineering design (IV). Model 2 uses experience teaching (IV). Model 3 uses both experience teaching engineering design (IV1) and experience teaching (IV2).

Model 3	B	t	Sig
Constant	8.377	34.524	0.000
Eng Des	0.066	3.566	0.001
Gen Teach	-0.035	-2.087	0.040

Appendix I: Informed Consent

North Carolina State University
INFORMED CONSENT FORM for RESEARCH

Title of Study/Repository: SETED validation and (eIRB number)>
Principal Investigator: Erik Ward, emward2@ncsu.edu
Faculty Point of Contact: Dr. Aaron Clark, aclarck@ncsu.edu

What are some general things you should know about research studies?
You are being asked to take part in a research study. Your participation in this study is voluntary. You have the right to be a part of this study, to choose not to participate and to stop participating at any time without penalty. The purpose of this research study is to gain a better understanding of engineering teaching tasks.

You are not guaranteed any personal benefits from being in this study. Research studies also may pose risks to those who participate. You may want to participate in this research because it will assist with the development process for an instrument to better understand engineering teaching tasks. You may not want to participate in this research because it may affect your confidence in your teaching ability.

In this consent form you will find specific details about the research in which you are being asked to participate. If you do not understand something in this form it is your right to ask the researcher for clarification or more information. A copy of this consent form will be provided to you. If at any time you have questions about your participation, do not hesitate to contact the researcher(s) named above or the NC State IRB office (contact information is noted below).

What is the purpose of this study?
The purpose of the study is to gather data to be used for analyzing a teaching engineering tasks survey.

Am I eligible to be a participant in this study?
There will be approximately 50-200 number of participants in this study.
In order to be a participant in this study you must at least 18 years old and currently be teaching in a classroom.
You cannot participate in this study if you are under 18, have not yet begun teaching in a classroom, or are student teaching.

What will happen if you take part in the study?
If you agree to participate in this study, you will be asked to do all of the following:

Complete the following survey.
Provide demographic information.

The total amount of time that you will be participating in this study is approximately 15 minutes

Risks and benefits

There are minimal risks associated with participation in this research. There are no direct benefits to your participation in the research. The indirect benefits are a different perspective on how you teach engineering may be gained. You may also be reflective of your own teaching.

Right to withdraw your participation

You can stop participating in this study at any time. In order to stop your participation, please contact the research or faculty contact above. You may also stop completing the survey at any time. If you choose to withdraw your consent and stop participating you can expect to have your data excluded from the study.

Confidentiality

The information in the study records will be kept confidential to the full extent allowed by law. Data will be stored securely on an NC State managed computer. Your IP address may be collected by the survey tool, but no identifiable information will be used by the research. Individual data with identifiable details removed may be made available to the public as required by a professional association, journal, or funding agency. To help maximize the benefits of your participation in this project, by further contributing to science and our community, your de-identified information will be stored for future research and may be shared with other people without additional consent from you.

Compensation
There is no compensation for participation in this study.

What if you have questions about this study?

If you have questions at any time about the study itself or the procedures implemented in this study, you may contact the researcher, Erik ward at emward2@ncsu.edu.

What if you have questions about your rights as a research participant?

If you feel you have not been treated according to the descriptions in this form, or your rights as a participant in research have been violated during the course of this project, you may contact the NC State IRB (institutional Review Board) Office via email at irb-director@ncsu.edu or via phone at 1.919.515.8754. You can also find out more information about research, why you would or would not want to be a research participant, questions to ask as a research participant, and more information about your rights by going to this website: http://go.ncsu.edu/research-participant

Consent To Participate

"I have read and understand the above information. I have received a copy of this form. I agree to participate in this study with the understanding that I may choose not to participate or to stop participating at any time without penalty or loss of benefits to which I am otherwise entitled."

Appendix J: Recruitment Email

Greetings,

My name is Erik Ward, I am a doctoral student at North Carolina State University. I am emailing looking for participants in my dissertation research. My research is the development of a skills appraisal inventory for various teaching tasks you may encounter or perform when teaching an engineering class or topics.

I am looking for participants that are over 18 and are in-service teachers teaching engineering or engineering design as part of their class or course load. If you could spare 10-20 minutes of your time to complete the survey using the link below it would be a great assistance to me and my research.

https://ncsu.qualtrics.com/jfe/form/SV_eDoo0XxJaCqzRBP

Thank you,

Erik Ward