Fall 12-2022

Investigating the Process of Consequential Validity with the Ambassador Questionnaire

Melissa Gayle Kuhn
Old Dominion University, stjohnmg@gmail.com

INVESTIGATING THE PROCESS OF CONSEQUENTIAL VALIDITY WITH THE AMBASSADOR QUESTIONNAIRE

by

Melissa Gayle Kuhn
B.A. May 2015, Early and Elementary Education, Virginia Commonwealth University
M.A., June 2015, Teaching PK-6, Virginia Commonwealth University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

EDUCATIONAL PSYCHOLOGY AND PROGRAM EVALUATION

OLD DOMINION UNIVERSITY
December 2022

Dissertation Committee:

Joanna Garner (Chair)

Shanan Chappell Moots (Member)

Linda Bol (Member)

ABSTRACT

INVESTIGATING THE PROCESS OF CONSEQUENTIAL VALIDITY WITH THE AMBASSADOR QUESTIONNAIRE

Melissa G. Kuhn
Old Dominion University, 2022
Chair: Dr. Joanna Garner

Validity in psychometrics refers to the degree to which evidence and theory supports the interpretations drawn from a test, and Messick's Contemporary Validity Theory (1994) includes several facets with well-established evidence collection methods. However, there is a lack of consensus on appropriate methods of evaluating the facet of consequential validity, which is the degree to which interpretation of scores could have consequences for test-takers. The primary objective of this study was to illustrate a method of identifying potential consequences of survey in the stage of manual development. This method was placed in the context of the Ambassador Questionnaire (AQ) used in Engineering Ambassador (EA) programs that aim to broaden participation in the discipline with outreach. Because this program focuses on equity and inclusion, a second objective was to determine if the AQ could include a subscale sensitive to equity perceptions. A mixed method was employed to address these objectives. Qualitative interviews informed both the capacity for interviews to solicit potential consequences of use and the utility of an equity perception subscale to EA chapter advisors. Quantitative analyses included factor analytic techniques. Existing AQ data was proportionately stratified for separate exploratory and confirmatory factor analyses to determine if controlling for demographic differences would reinforce the AQ's existing factor structure that omits pilot items sensitive to perceptions of equity differences.

Ten advisors were interviewed and a total of eleven hours of interview data was collected. Several themes related to consequences of use emerged from these interviews, including use of the AQ to track student growth, potential influences on students as a result of taking the AQ, programmatic changes that could be made with AQ data, and sharing of AQ data with interested parties. These findings suggest that interviewing stakeholders during the development of an instrument's manual can provide developers with foreseeable consequences. Integration of advisor feedback on items sensitive to perceptions of equity differences with quantitative findings that reinforce the AQ's existing factor structure also suggests that a scale sensitive to said perceptions would be best served by an additional subscale with added items or an independent equity perceptions scale.

This dissertation is dedicated to my family and friends. Thank you to my parents, brother, grandmother, and in-laws for believing in me: thank you Nick for always loving and taking care of me. To my friends, I need to thank Angel for supporting and pushing me towards this path, Hong for inspiring and going through this with me, Abigail for always listening and grounding me, Mimi for proving Premack's Principle in the most unexpected ways. Your perspectives, compassion, and company made a difference.

At the foundation of inequity is colonialism: this dissertation was developed and written on the lands near the Powhatan Confederacy's Chesapeake and Nansemond tribes, which were driven from their homes during the early colonization of Eastern Virginia through assimilation, encroachment, and war despite treaties of land protection promised to them throughout the 17th century. Later, Virginia's Racial Integrity Act of 1924 meant many descendants of the Powhatan Confederacy also lost records of lineage.
If we wish to move the needle, we first must know the history of where it is stuck.

ACKNOWLEDGEMENTS

There are many people who have contributed to the successful completion of this dissertation. First, I must thank my committee members – Dr. Joanna Garner, Dr. Shanan Chappell Moots, and Dr. Linda Bol – for their guidance, feedback, and mentorship throughout this process.

Thank you, Dr. Sally Wei, for your assistance with the Engineering Ambassadors Network, including providing resources and contacts for potential participants in this study. I also must thank Dr. Melissa Quackenbush, who provided an external audit of the qualitative findings of this study, providing valuable insight and perspective while doing so.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND OVERVIEW

Problem and Aims

Validity in psychometrics is a complex issue; the most commonly understood meaning of validity in measurement is the degree to which an assessment tool measures what it intends to measure (Furr & Bacharach, 2014). However, this definition is an oversimplification. More accurately, validity refers not to the instrument itself, but the degree to which evidence and theory supports the interpretations one draws from a test (Furr & Bacharach, 2014). It is not definite and is unquestionably an ongoing process. The six major facets of contemporary validity according to Messick (1994) include content validity, substantive validity, structural validity, generalizable validity, external validity, and consequential validity. Of these six, this dissertation study endeavored to investigate how the instrument development process can further integrate strategies for soliciting consequences of use towards the end of the iterative cycle that instruments so often go through in development and validation. The first five of the facets identified by Messick (1994) are commonly accepted in psychometric research and have identified, standardized means of providing evidence (Furr & Bacharach, 2014). The consequences of use of an assessment, however, have remained contentious since Messick's inception of the theory and lacks consensus on how consequences can be evaluated, let alone when one should consider consequences of use. Perhaps one of the greatest arguments against consideration of consequential validity at any stage of the instrument development and validation process is the extent to which identified potential consequences could influence score inferences (Cizek, 2020). However, this argument assumes that score inferences will be made without consequence-informed error from those who are using the scores, such as using the scores to

draw conclusions or make decisions that are incongruent with the instrument's design. It further assumes that there is an absence of construct-irrelevant variance that could be informed by said consequences, as is often the argument against high-stakes college entrance examinations popularly believed to exacerbate socioeconomic differences in subgroup access to higher education. Consequential validity, for the purpose of this dissertation study, should be understood as a component of construct validity as well as validity evidence that identifies the intended and unintended consequences of use of the scores derived from an assessment, adapted from Messick's Contemporary Validity theory.

Approaching the Validity Issue

There are plentiful sources of validity evidence that researchers can argue for over the course of an instrument's development, and best practices and accepted methods of obtaining validity evidence are fairly standardized. A few examples of these methods include careful construct definition, inductive and deductive item generation, expert panel reviews, dimension reduction techniques, and concurrent validity comparisons (Boateng et al., 2018). Notably, these methods combine qualitative and quantitative methodological strategies at different stages to support the validity of test interpretations. However, as an instrument moves towards finalization, there is a marked absence of qualitative strategy; revisiting expert panel reviews with subject matter experts does not typically occur after item reduction and finalization of an instrument's subscales. Nevertheless, consequential validity is often considered an important part of construct validity; there is just little consensus on how to best identify and explain what the consequences of use of an instrument may be (Furr & Bacharach, 2014). Construct validity broadly could also be investigated again with qualitative strategies towards the end of the instrument development process in the same way that expert panel reviews are used at an earlier

stage, prior to actual data collection. As instruments often change in constructs and quantity of items over the course of the development process, it is curious that standard measure development protocols do not include consultation with various stakeholders related to an instrument, especially those who utilize it.

The Impetus for Developing the AQ

The Ambassador Questionnaire (AQ) is an omnibus assessment tool that has undergone development and an initial validation study (Kuhn et al., 2022). While ample evidence exists to support the ability of engineering outreach programs such as engineering ambassadors to act as an effective recruiting arm of their program and broaden K-12 students' interest in engineering (Smaill, 2010; Moskal & Skokan, 2011; Nadelson & Callahan, 2011), research on the undergraduate ambassadors who participate in outreach activities is still a growing field in need of tools that can measure and quantify how outreach is perceived by the ambassadors. Qualitative research by Garner and colleagues (Garner, et al. 2018) suggested that ambassadorship impact manifests in ways that cohere to form a role identity for the engineering ambassador that includes particular engineering and outreach related self-perceptions, purposes and goals, beliefs, and possibilities for action. The work provided a theoretical basis for much needed instrument development: an initial 2018 survey of 30 chapters of engineering ambassador programs found that less than 10% routinely assessed the impact of their programs on the participating undergraduate students (Garner et. al., 2019) and in focus groups in the Spring of 2018, a consensus arose among program advisors of the need for a validated tool that could assess the influence of participation on the ambassadors themselves. The developed AQ was theorized to involve several constructs found in EAs with the Dynamic Systems Model of Role Identity (DSMRI; Kaplan & Garner, 2017). The AQ includes five validated subscales at present:

Academic Confidence, Engineering Beliefs, Professional Skills, Social Supports, and Ambassadorship, which map onto concepts from the DSMRI such as ontological and epistemological beliefs (Engineering Beliefs), purpose and goals (Academic Confidence, Ambassadorship), self-perceptions and definitions (Professional Skills, Social Supports), and perceived action possibilities (qualitative, open-ended items) (Kaplan & Garner, 2017; Kuhn et al., 2022). Prior qualitative work with the DSMRI found all four major components of the model were present in an EA role identity (Garner et al., 2018). The instrument was validated with traditional item analyses processes (Kuhn et al., 2022), but had yet to be inspected for consequences of use. In response, the primary aim and contribution of this dissertation study was an investigation into the construct and consequential validity of this validated measure of ambassadorship impact.

Application to Engineering Education

This study endeavored to investigate this gap in emphasis on construct and consequential validity towards the conclusion of the instrument development and validation process with an omnibus assessment tool developed to assess components of role identity development of undergraduate students participating in outreach activities, known as Engineering Ambassadors (EAs). Conceived as a tool to assist EA program advisors in the systematic evaluation of their programs, the AQ is a unique tool designed for both use in both research and practice. As an instrument designed for researcher and practitioner use in informal learning settings, the AQ addresses a gap in the available tools for EA chapter advisors. As this uniquely situated instrument is available and has prior research, potential consequences of use of this instrument can be solicited from the intended users.

Despite the lack of available tools for use with research and practice with EAs, much learning and personal growth occurs in outreach programs that warrants further investigation. In a systematic review of 51 studies investigating the impact of outreach activities on undergraduate engineering students, only two studies explicitly cite use of validated instruments in the research design (Garner, Thole, & Alley, 2020). Using researcher-generated measures, 65% of reviewed studies reported impacts on undergraduate students' professional confidence and technical skills in communication and presentations (Garner et al., 2020). Prior to the studies identified in this systematic review, research involving engineering outreach programs more often focused on student recruitment outcomes, as student ambassadors are an asset to most engineering programs for generating prospective student interest.

Outside of research affiliated with the national Engineering Ambassadors Network (EAN), assessment of the growth and development of the participating undergraduates is still limited. However, a broader body of research outside of outreach programs suggest that there are several positive benefits to participating in service roles, such as ambassadorship: participating in engineering service projects to others or one's institution lends itself towards increased engagement and persistence in the discipline. Engineering outreach and service programs may also reduce attrition among engineering students, particularly isolated female students and students of color (Litchfield & Javernick-Will, 2015) and aid engineering students in professional socialization and the development of a career-based identity (Laursen et al., 2012). Ambassador group advisors play a key role in monitoring and improving the experiences of ambassadors and may wish to use a survey measure of ambassadorship impact to better understand their ambassador cohorts and provide professional development experiences that

maximize their time in the program. Therefore, a second, exploratory aim of this dissertation study was to investigate the measure's perceived utility for practitioners in the field.

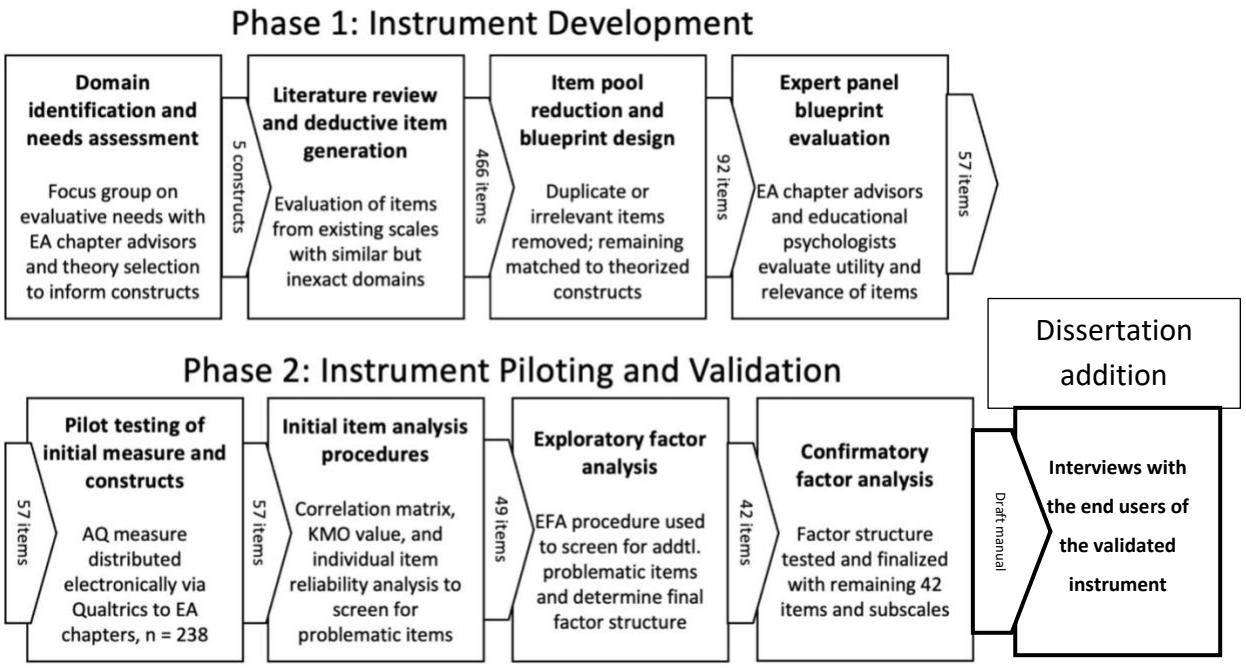
Prior Research with the AQ

The prior work relevant to this dissertation study grew from the qualitative work by Garner et al. (2018) to involve discussion with EA chapter advisors, wherein the need for a valid and reliable tool for assessing students in their programs was revealed. Consequently, existing measures relevant to engineering education, engineering beliefs, professional skills, and interpersonal skills in a university setting were gathered and analyzed for useful content (Kuhn et al., 2019). Measures that tapped into relevant contexts include the Longitudinal Assessment of Engineering Self-Efficacy (LAESE), the Project to Assess Climate in Engineering (PACE) survey, the Laanan Transfer Students Questionnaire (L-TSQ), the National Women in Engineering Programs and Advocates Network (WEPAN) pilot climate survey, the Academy of Engineering's *Changing the Conversation* pilot survey, the Assessing Women and Men in Engineering (AWE) retention survey resources, and Engineering Ambassadors Network chapter exit interview protocols. All measure items were gathered under free use from the publications of the measures in question (L-TSQ, PACE, WEPAN, LAESE, *Changing the Conversation*), stated free use for the purposes of researchers interested in measuring qualities of engineering students (AWE), or via direct contact with persons owning the measures (EAN exit interviews). The prior work aimed to develop an instrument for use by researchers and practitioners and resulted in a 41-item scale with five subscales that include constructs that support the development of an ambassador role identity (Kuhn et al., 2022). Standard instrument development practices were followed, such as the use of a draft blueprint and expert panel review prior to pilot testing of the instrument, iterative periods of data collection, and exploratory and confirmatory factor analyses

were conducted to identify and verify the instrument's factor structure (Boateng et al., 2018). By most standards and practices in instrument development, the validated version of the AQ (Kuhn et al., 2022) presents sufficient evidence for both reliability and construct validity. However, this standard process for instrument development was not able to provide conclusive evidence for items pertaining to perceived equity differences and has not included input from end users who may provide insight on the consequential validity of the instrument.

Prior validation efforts followed a traditional and standard approach to the instrument development process, as seen in Figure 1 (Boateng et al., 2018; Kuhn et al., 2022) but some questions persisted regarding variables that may influence response patterns and the potential to explore a process for gathering consequential validity evidence.

Figure 1. ***Modified Instrument Development Procedure*** **(adapted from Kuhn et al., 2022)**



Prior work revealed that there was potential for the AQ's scores to be sensitive to demographic variables arising from differential response patterns for certain questions (Kuhn et

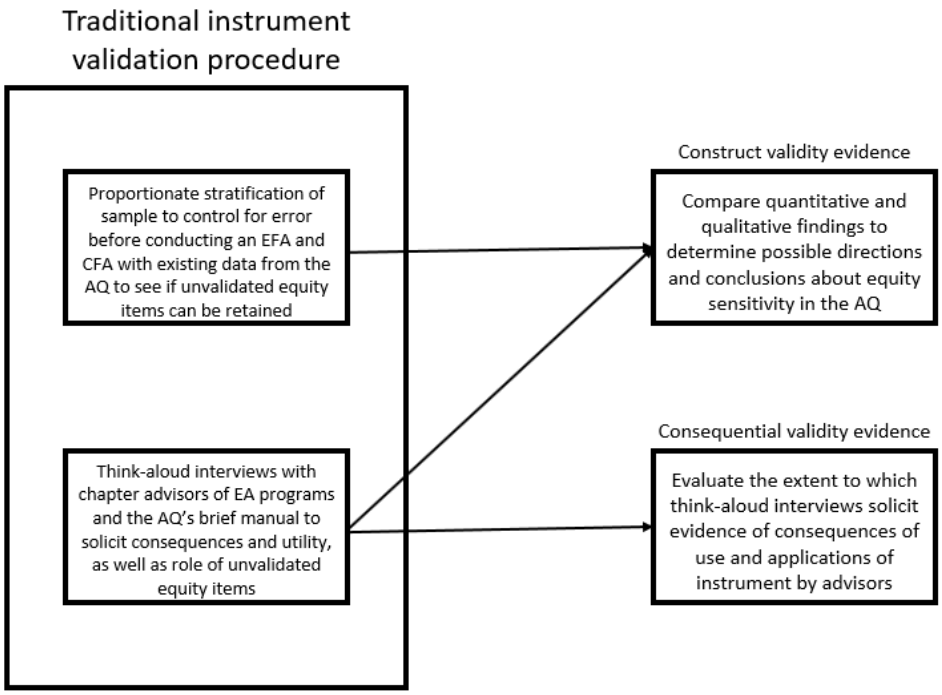
al., 2019) but these items were not able to be retained in the final factor structure (Kuhn et al., 2022). These items included the following statements for a five-point, Likert-type agreement scale: “I am confident I will be treated fairly on the job in engineering,” and “I am confident I will be given the same opportunities for pay raises and promotions as my fellow workers if I enter engineering.” Proportionately stratifying data by demographic identifiers (Salkind, 2010) would help ensure that there was almost no variance in demographics between the split pool of data for the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) procedures. Aside from the removed items that demonstrated the potential for detecting gender differences in responses, interim investigations of AQ data with chi square tests between the two randomly split subgroups for the EFA and CFA revealed that there were some response pattern differences by years as an ambassador (Kuhn, unpublished pilot study). The random split that was used for the validation study did not explicitly attempt to control for potential subgroup differences (Kuhn et al., 2022). This points to investigating the AQ data anew with a proportionate stratification strategy to control for differences and inequalities between the measure’s EFA and CFA analysis groups. Furthermore, the items in the validation study were not reviewed by the EA chapter advisors. Considering the AQ was developed from a need expressed by chapter advisors (Garner et al., 2018) this presented an opportunity to establish construct and the potential consequential validity towards the end of the instrument development process.

Project Scope

This dissertation study was an embedded design mixed-methods approach (Creswell & Plano Clark, 2007) to address a gap in the instrument development and validation process wherein the end users of an instrument are not typically consulted to gather consequential and

construct validity evidence. In addition, quantitative evidence from a proportionate stratification and additional exploratory and confirmatory factor analysis can co-inform additional evidence of construct validity. Semi-structured interviews informed by Messick's contemporary validity theory (1994) were used to review the validated AQ and its practitioner manual with EA chapter advisors. This project concurrently investigated the potential to retain items that may be sensitive to differences in equity perceptions by subgroup using proportionate stratification. During the resulting EFA and CFA process, standard procedures for both factor analyses were implemented. The model of this embedded design for this dissertation study is included in Figure 1, pictured below. The study includes two new steps within the traditional instrument validation procedure, namely a new strategy for assigning participants to EFA and CFA procedures and conducting stakeholder interviews during instrument manual development. These additional steps will allow for added construct validity evidence as well as evidence of foreseeable consequences of use.

Figure 2. ***Study design***



Because the AQ was developed with practitioners and researchers in mind, it is well-situated for testing an interview method with individuals who often come with a wide range of relevant expertise. Therefore, this study aimed to test a new strategy for soliciting construct validity evidence regarding an identified gap with the AQ as well as consequential validity evidence regarding the stakeholders' imagined uses of the instrument. Considering past research with the AQ included items relevant to perceptions of equity in engineering (Kuhn et al., 2019) that were also sensitive to gender differences, this study also investigated whether items previously removed from the scale should be retained or separated into their own subscale. The research was guided by two qualitative research questions and one quantitative research question.

The qualitative research questions are: 1) In what ways does the AQ have practical, consequential use for chapter advisors? 2) What recommendations do EA program advisors make with regards to improving the AQ?

The quantitative research question is: 1) How can the construct validity of the AQ be enhanced through proportionate stratification of participants into separate factor analysis groups?

CHAPTER II

REVIEW OF LITERATURE

The process of developing a survey instrument for social science research includes many recommended best practices (Boateng et al., 2018). Following the process outlined by Boateng et al., the process starts with identification of the domain and item generation (2018). Item generation can take place inductively with focus groups and interviews, or deductively through a review of existing measures and literature to identify potential items that will encapsulate the defined construct (Boateng et al., 2018). After an initial blueprint has been established, it is recommended to evaluate the items as they pertain to their respective domains with expert review and evaluation, and/or target population evaluation (Boateng et al., 2018). Once a survey has been pared down to appropriate and necessary items to evaluate the construct in question, the scale should be piloted in the target population, and care should be taken to ensure that an adequate sample size is achieved (Boateng et al., 2018). After pilot data has been collected, item reduction procedures should be used to ensure the scale is parsimonious, such as using inter-item and item-total correlations and ensuring that the sample of responses is adequate for each item with little missing data (Boateng et al., 2018). This leads to the initial factor analysis to determine the optional number of factors for the set of items, which is followed by the final phase of scale validation (Boateng et al., 2018). In the final stage of developing a scale, it is important to test the dimensionality of the latent factor structure with methods such as confirmatory factor analysis, as well as establish reliability of the instrument (Boateng et al., 2018). The intent of these stages is to assess an instrument at various stages of its development in a rigorous way that contributes to the evidence for the validity of the instrument. The final stages often include looking into other sources of validity evidence, such as assessing predictive

validity with future outcomes and concurrent validity with other instruments for criterion validity, and construct validity through examining convergent validity with different assessment approaches and discriminant validity with a different concept (Boateng et al., 2018). Combined, these various steps and common practices in psychometric survey development intend to provide evidence for the validity of an instrument at multiple stages, gradually shifting from more qualitative approaches of gathering data to more quantitative methods of assessment and analysis. However, it is important to consider that instruments themselves are not what validity evidence argues for. The present work examined this process and considered concepts of validity theory, especially how validity can be enhanced with additional qualitative evidence gathered towards the end of the instrument development and validation process.

Theoretical Framework for the Qualitative Research Questions: Validity and Measure Development

Validity is an important component of the development and finalization of any instrument designed to quantify psychological constructs. Without evidence for validity, there is no way to know for certain that an instrument is measuring what it was designed to capture. In some cases, a lack of evidence for the validity of an instrument can have dire consequences, such as a hypothetical poorly designed medical school entrance exam that admits unqualified students. More commonly, a lack of evidence for the validity of an instrument used in a study is a significant weakness and threat to potential findings in psychological and educational research. The simplest way to describe validity is described by Furr and Bacharach as “the degree to which a test measures what it is supposed to measure” (2014), which is also described by the authors as a major simplification that does not quite capture the full complexity of the idea. Validity can never be definitively proved, it can only be argued for with evidence and theory to support the

interpretations of the test scores (Furr & Bacharach, 2014). What is ultimately quantified as valid or invalid with the available evidence for validity is the interpretations of the test outcome, not the assessment. Commonly accepted forms of evidence can include content validity, face validity, construct validity, convergent validity, and discriminant validity (Furr & Bachrach, 2014). However, an important component of Messick's Contemporary Validity Theory (1994) has remained contentious for more than two decades: consequential validity (Furr & Bacharach, 2014) with long-standing debate over consequential validity's role in the measure development process persisting to this day. To evaluate consequential validity, few strategies are offered by Furr and Bacharach; the authors merely highlight the scientific debate over consequential validity's inclusion as an area of concern in psychometrics. However, consequential validity is highlighted for its association with construct validity; one of the primary means of asserting construct validity is with expert panel review of a test blueprint early in the measurement process, item reduction and extraction of factors midway through the process, tests of dimensionality, and validation with convergent, criterion, and discriminant evidence (Boateng et al., 2018). Notably, construct validity is largely quantitative in approach in Boateng et. al.'s final measure evaluation stage (2018); there is little mention of using panel reviews or cognitive interviews once the measure's pilot data has been collected.

Messick's Perspective on Validity

Validity, according to Messick (1988) always refers to the extent to which empirical evidence and theoretical rationale support the adequacy and appropriateness of the conclusions one draws from the outcomes of a given test score. This conception of contemporary validity involved six major components: content validity, substantive validity, structural validity, generalizability, external validity, and consequential validity. In the time that has passed since

Messick presented a concrete definition of validity and a framework for evaluating the validity of psychological instruments, researchers have presented theoretical evidence and rationales for the validity of their measures, including testing concurrent validity of instruments and the structural validity of instruments through exploratory and confirmatory factor analysis (Flake et al., 2017).

However, many developed and published instruments have focused less on an important component that Messick emphasized (1994): consequences of use, or consequential validity. Messick's work in the development of Contemporary Validity Theory argues that social consequences of the interpretations and use of a test should be as important as the developmental steps taken to prove that a test is able to measure what it is designed to (1994). Evaluating how well a test or survey performs at the task it is designed for should have both the predicted and actual social consequences considered; if potential consequences interfere with the purpose of the test or go against commonly held social values, a threat to validity may exist (Messick, 1994). Consequential validity is perhaps best understood as a component of construct validity, as construct validity overall is "the meaningfulness or trustworthy interpretability of the test scores and their action implications" (Messick, 1994, p.15).

While the other facets of contemporary validity are doubtlessly important, there is little debate over the merits and methods of attaining evidence for the first five forms of validity. Messick's interpretation and suggestions for the other five sources of validity are well established. Content validity focuses on content relevance, representativeness, and technical quality (Messick, 1994), which can be attained with expert panel review. Substantive validity involves the actual processes aligned with the construct's tasks and the precision of the instrument (Messick, 1994). Structural validity examines the relationship between scoring structures to the content domain (Messick, 1994) which is typically achieved with factor analysis

procedures. Generalizability looks to examine how score properties relate to the broader population, or across different populations (Messick, 1994) which is attainable with modification and retesting of instruments with different subgroups. External validity includes methods of convergent and discriminant validity; these are easily attained from multi-trait and multi-method comparisons (Messick, 1994) such as comparing the outcomes of a test with similar constructs.

All five of these components have commonly accepted means of testing and reporting in the psychometric community; remaining challenges with Messick's conception of validity are found in consequential validity (Furr & Bachrach, 2014). However, Messick argues that evaluating the consequences of use of a test should be especially important with respect to potential for bias; if construct underrepresentation or construct-irrelevant variance are present in an assessment, these could lead to scoring that does not adequately reflect a participant's performance or ability (Messick, 1994). Therefore, when evaluating the consequential validity of an assessment, one may wonder why evaluation from the end users that an instrument is designed for is not commonly included as a step to ensure additional validity of an assessment after pilot data collection and statistical analyses have been conducted with the pilot data. Considering that many instruments start with a pool of items that is gradually reduced throughout the development process, the instrument may change in consequences of use and constructs by the end of an examination of the instrument's psychometric properties.

Perspectives on Validity in Educational and Psychological Measurement

While Messick's perspectives on validity will be the primary underlying theory with which to examine consequential validity and evaluate the consequences of use of the AQ, there are other perspectives on validity in educational and psychological measurement to consider. Marsh (1987) has highlighted the difficulties and conundrums that may arise when attempting to

quantify learning and teaching quality in higher education, specifically comparing optimal performance exam scores of students to typical performance evaluations of teaching assistants in different college sections, arguing that a teaching assistant who pushes their students harder may earn lower evaluations from the same students who earn the most impressive exam scores. Marsh also discusses that, while ideal to improve the standards of psychometric evaluation, there are rarely longstanding theories in the field and multiple sources of data typically need to be collected to be able to draw any reasonable conclusions. However, the works of Marsh in respect to validity typically correspond to improving the validity of methods of teacher evaluation and student learning (Marsh, 1987; Marsh & Roche, 1997) or psychological constructs of self-concepts and motivation (Marsh et al., 1983; Marsh et al., 2013) rather than career related identity and self-perceptions, which is the focus of the AQ.

Other researchers have highlighted the number of different methods and processes with which researchers can collect evidence for an instrument's validity (Flake & Fried, 2020). As the ways in which measures can be developed and interpreted vary broadly, careful judgement and transparent reporting is required to ensure that the interpretations one may draw from the measure are valid (Flake & Fried, 2020). Flake and Fried further offer a framework for promoting the transparent reporting of measurement practices, which in turn promotes the validity. The framework includes identifying and defining the construct, justifying the measure selection, and presenting any existing validity evidence, describing how the measure operationalizes the construct, describing how the measure is quantified, describing if/how a scale was modified, and explaining how or why an existing measure may have not been used.

However, several issues persist with consistency of reporting of these ideas to support the validity of researcher-developed instruments. For instance, when instruments are used in

psychological research, authors of publications sometimes provide only limited psychometric evidence or description of the properties of the measure and do not question important psychometric properties like reliability and validity (Flake et al., 2017). While standard, best practices in scale development to aid evidence of validity are prevalent and broadly discussed in Flake and colleagues' work, there are still challenges with ensuring that standard practices are widely known. Considering the contentious discussion around consequential validity illustrated by Furr & Bachrach (2014) and the issues that Flake and colleagues highlight about the reporting of instruments to ensure that the validity evidence is clearly presented, it would follow that there are still issues with reporting validity in contemporary measurement. Indeed, measures often move in cyclical processes of development, validation, repurposing, and revalidation; validity is never absolute with measures, and instead supports the interpretations ascribed to an instrument. This is partly why some of Flake and colleagues' research focuses on the reporting of validity and reliability evidence of instruments used after they have previously been subjected to validation procedures (Flake et al., 2017). Thus, there seems to be room for improvement in how validity evidence is ascertained due to the nebulous nature of reporting of validity, and ongoing debate over how to interpret the consequences of use component of Messick's theory of validity.

Theoretical Framework for the AQ and the Quantitative Research Question

What Kuhn et al. (2022) hoped to better understand while quantifying the qualities of EAs is the students' own concept of what it means for them to be an EA. To this end, an identity-based theory was used in the development of the AQ. EA programs typically include a large number of historically underrepresented engineering students, whose reasons for choosing engineering and outreach have not been well described in the literature. EA program participation may stem from students' desire to "promote the next generation" and model to K12

students during outreach activities that an engineer can look like a wide range of persons from diverse backgrounds. Indeed, socially engaged engineers tend to be more female and more culturally diverse than the norm, with service organizations such as Engineers Without Borders attracting far more female engineers and engineers of color with motivations towards social good (Litchfield & Javernick-Will, 2015). It may also be theorized from Litchfield and Javernick-Will's work involving service-oriented engineers that EAs likely have their own unique role identity as an EA that explains their cognitions and motivations, as participants specifically separate themselves out from engineering stereotypes as "being social, not introverted, musical, or the weird… social justice person" (Litchfield and Javernick-Will, 2015, p. 400). Against this backdrop, the unique identity of engineering students and its capacity to explain why students enter engineering in the first place remains a well-documented phenomenon (Godwin et al., 2016). Survey work from Godwin et al. (2016) also reaffirmed the association between historically underrepresented groups in engineering and commitments to prosocial impacts of engineering, the presentation of which is a central tenet of EA programming. With these observations from the current body of engineering education and outreach literature in mind, characteristic features of an EA's role identity as an ambassador may exist and potentially change throughout their time in the program. These include their beliefs, motivations, self-perceptions, and possibilities for future action, which can be tied together using the Dynamic Systems Model of Role Identity (DSMRI; Kaplan & Garner, 2017) to inform a measure that investigates the qualities of EAs followed.

Foundational qualitative research ahead of the development of the AQ applied the DSMRI to the observation of professional and interpersonal growth in EA students (Garner et al., 2018). As its name implies, the DSMRI is a complex and dynamic systems model through which

one can view social role as a central unit of analysis (Kaplan & Garner, 2017). Within specific social contexts, a role identity is shaped by four interacting and codependent facets: one's ontological and epistemological beliefs, one's purposes and goals, one's self-perceptions and definitions, and one's perceived action possibilities (Kaplan & Garner, 2017; Garner et al., 2018). A role identity can be something that is extremely specific and contextual, such as two hours of membership in an audience during a symphony orchestra (and knowledge that such a role means one reserves clapping until the end of the entire piece, not just specific movements) or long-lasting and occupational, such as taking the role of a veteran teacher in a K12 classroom (and knowledge of the role-specific processes associated with the identity of a veteran teacher in K12 settings, such as best practices for parent-teacher conferencing). EAs are likely to have multiple role identities that coexist within them, including (but not limited to) a role identity as an engineering student, a separate role identity as an ambassador for the field of engineering, and another role identity as a future engineer. On a multi-year scale, an EA-specific role identity could motivate behaviors such as engaging in continued outreach activities, increasing confidence in the use of interpersonal and professional skills to communicate engineering concepts to K12 audiences, and becoming highly involved in the governance structure that the local EA program maintains.

Garner et al. (2018) found that prior role identities could motivate students to become an EA in the first place, as well provide evidence that the professional development aspect of EA programs help shape and solidify a student EA's identity not just as an engineering student, but as an engineering student who is an ambassador. A measure that attempts to capture the role identity of the EA can be easily applied to three of the four model components, with different domain-specific investigations within one of the four: epistemological beliefs about the nature of

engineering (“Engineering Beliefs”) self-perceptions and definitions in an engineering academic domain (“Academic Confidence”), self-perceptions and definitions in a social domain (“Social Supports”), self-perceptions and definitions in a professional competence domain (“Professional Skills”), and goals and purposes as an EA (“Ambassadorship”). Perceived action possibilities are perhaps best addressed with qualitative, open-response items as opposed to Likert-type, quantitative, agreement-based subscales. Therefore, open response items to address perceived action possibilities (such as intent to engage in outreach after graduation) could be included in a separate section of the final draft of the measure. Many of the resources currently used by EA program advisors include open-ended response questions that would elicit descriptions of an EA’s perceived action possibilities (University of Nebraska, 2018).

This theoretically grounded, qualitative research served as the initial framework for a quantitative measure of ambassadorship – the AQ. When the AQ was developed, a deductive item generation process was used and relevant literature in engineering education and measurement was sampled, reduced, and matched to the above identity-related constructs (Kuhn et al., 2022). With an initial pool of 466 sampled items reduced to a blueprint of 92 potential pilot items that matched the desired constructs, an expert panel was used to support construct validity (Boateng et. al., 2018). The panel included four EA chapter advisors and one educational psychologist familiar with both the theoretical framework and EA programs (Kuhn et al., 2022). The blueprint asked the expert panel to rate the importance of the items to the purpose of the measure on a Likert-type rating. After the removal of redundant or irrelevant items, the resulting piloted omnibus measure of EA role identity consisted of 57 items, which was further reduced to 41 after item reduction protocols and factor structure analysis to establish evidence for construct

validity and the alignment of the factors with the theorized identity-related components. (Kuhn et al., 2022).

Context: Engineering Ambassadorship

Ambassadorship in higher education is typically a form of outreach conducted by undergraduate students using several different approaches, including presenting outreach presentations to K12 school groups, providing campus tours to prospective students, providing mentorship and tutoring, engaging in hands-on activities with K12 students, and participating in prospective donor events (Garner et al., 2019). Oftentimes, groups are intended to broaden the recruiting efforts of a specific institutional program, such as a college or school of engineering (Hatzell et al., 2013) or a school of pharmacy (Kwiatkowski et al., 2019); however, ambassadors can be any undergraduate student engaged in some of outreach and recruitment on behalf of their institution (Green, 2018). Broadening the recruitment pipeline in K12 has long been a major focus of engineering programs (Orsak, 2003; Molina-Guado et al., 2010) and broadening participation from underrepresented groups in engineering has been highlighted as the central focus of many Engineering Ambassador (EA) groups (Hatzell et al., 2013). To promote these efforts with modeling, some programs focus their recruitment on undergraduates from underrepresented groups, such as women (Hatzell et al., 2013). Women in engineering also tend to have higher rates of participation in volunteerism and feel a greater sense of social responsibility (Canney & Beilefeldt, 2015), so an overrepresentation of women in engineering outreach is unsurprising. Often overlooked in earlier research on EA programs were the outcomes and benefits for the undergraduate students, while said outreach programs proved effective at sparking interest and motivation in K12 students for entering engineering or other related science, technology, engineering, and mathematics (STEM) fields (Habash & Suurtamm,

2010). Additional evidence has found that the presence of ambassador programs improves and helps maintain enrollment through creating a sense of community (Bates et al., 2014).

More recently, research has investigated the different ways that engineering students personally benefit from their participation in outreach or other kinds of volunteer activities (Anagnos et al., 2014; Garner et al., 2018; Huff et al., 2016; Laursen et al., 2012). However, until recently, there was no systemic, standardized, validated means of evaluating impacts on the participating undergraduates (Garner et al., 2019), which was part of what the construction of the AQ was intended to address.

The Nature of EA Programs

The nature of any given EA program from institution to institution can vary widely. EA programs sometimes operate like clubs, with self-organization and self-governance, including elected roles for participating members (University of Connecticut, 2022). Some programs are entirely student-run, while others have frequent oversight by faculty advisors (Hatzell et al., 2013). Some programs can have as few as seven ambassadors, while some have 100 or more (Haas et al., 2016). Broadly speaking, there is a wide range of membership requirements and organizational structures in EA programs. Only a subset of EA programs are affiliated with a national organization, the Engineering Ambassadors Network (EAN) (Haas et al., 2016). To join the EAN, an EA program must agree to the following standards: 1) they are to support a diverse group of ambassadors, with a focus on recruitment of underrepresented students of color and of diverse gender identities; 2) ambassadors are trained intentionally and purposefully in communication skills, and professional development in communication and leadership are central to network expectations; 3) ambassadors connect students to engineering, and outreach to K12 students with messaging from *Changing the Conversation* (National Academy of

Engineering, 2008) is used to accomplish this connection; 4) the program participates in annual network trainings and shares resources with other schools; and 5) the program has a faculty or staff advisor who oversees the program (Haas, 2016).

A common experience shared by EAs is the public-facing nature of participation in these programs. EAs in all programs, regardless of network affiliation, must develop interpersonal and professional skills to be able to engage with a diverse range of audiences in a diverse range of activities. Because of this crucial, core component of what EAs do for the engineering programs they serve, a core focus of the EAN as previously mentioned is to provide affiliated programs with the professional development opportunities that would facilitate marked gains in these communication and leadership skills (Haas et al., 2016). Public speaking and advanced instruction in communication skills are required for continued participation in EA programs, and implementation of specific messaging from *Changing the Conversation* that is found to appeal to historically underrepresented populations help ensure the effectiveness of these programs that engage in intentional training and professional development (Thole et al., 2013). Prior studies have looked specifically at what has been described as soft skills in EA programs: teamwork, communication, management, and leadership (Anagnos et al., 2014). Other studies still have looked at identity development of EAs as it pertains to motivations for outreach participation, including the goals, self-perceptions, beliefs, and action possibilities (Garner et al., 2018). More broadly, ambassadorship roles in other disciplines have been found to promote similar growth in soft skills, such as communication, introspection, and interpersonal respect (Kwiatkowski, 2019). Furthermore, students engaged in ambassadorship with an objective of broadening participation in postsecondary education may experience increased sense of belonging and retention in their own educational attainment (Green, 2019). Large quantitative studies with undergraduates

participating in some form of volunteer service have established a connection between academic development and life skill outcomes for some time (Astin & Sax, 1998). With the body of research on EA programs growing and demonstrating and impact on the students, prior work endeavored to develop a psychometrically valid and reliable quantitative tool these programs could use (Kuhn et al., 2019; Kuhn, et al., 2022). With an accumulation of empirical evidence that suggests that ambassadorship can lend itself towards growth for the students who facilitate outreach, a greater understanding of the core messaging focused on in EA programs in particular is necessary to establish what may make EAs unique from any other kind of ambassador.

Messaging in EA Programs

Messaging around engineering is one of the core components of EA training, and in this context, it should be understood as the way that one speaks about a given field; it is statements, associations, assertions, and ideas about the field that shape the way others perceive and think about it. The degree of agreement with prosocial messaging found on the AQ asserted by ambassadors taking the survey provides a snapshot of training effectiveness and buy-in. Over the last few decades, millions of dollars have been spent to try to improve public understanding of engineering and the role it plays in society, yet perceptions of the field tend to emphasize the career opportunities for engineers and the necessity for strengths in science and mathematics curriculum (National Academy of Engineering, 2008). As a result, many students have limited understandings of engineering; while they often comprehend that engineers design and build things, other associations with the field include believing that the work is solitary and sedentary, the work requires unattainable intelligence, and engineering makes good money (National Academy of Engineering, 2008). However, when studying different messaging around engineering, the National Academy of Engineering's *Changing the Conversation* study found

that certain messages around engineering that emphasize the field's role and contributions in society appeal to groups of students most critically underrepresented in the field, including women of color (2008). These messages in *Changing the Conversation* include statements such as "Engineers make a world of difference" and "Engineering is essential for our health, happiness, and safety" (p. 9). The report found that public perception of engineering was generally positive, despite perceptions from within engineering of how the field is perceived; as a career path, engineering is perhaps not fully understood, but is still respected and looked upon favorably by students and parents.

Drawing upon comparisons to medicine, the *Changing the Conversation* report also highlights how a strong emphasis on math and science challenges in the field can be discouraging and perceived as a barrier by students who may otherwise show interest in the field; medicine also possesses challenging curriculum in math and sciences, yet the emphasis on the curriculum in medicine is not as prevalent compared to the emphasis on relieving human suffering and bettering the lives of others (National Academy of Engineering, 2008). In total, for specific messages were identified by the *Changing the Conversation* report as potential cornerstones that would broaden participation in engineering and improve public perceptions of the field: 1) engineers make a world of difference; 2) engineers are creative problem solvers; 3) engineers help shape the future; and 4) engineering is essential to our health, happiness, and safety (p. 12).

One of the main recommendations from the National Academy of Engineering to facilitate changes in messaging around engineering was to use outreach in particular with K12 students. At its inception at Pennsylvania State University in 2009, this recommendation was taken up in the engineering program's new EA chapter (Thole et al., 2013). The next year, grant

funding helped expand the EA program at Pennsylvania State University into a network of four northeastern universities who shared a common aim of implementing the *Changing the Conversation* messages (Hatzell et al., 2013). At present, the EAN includes over 30 institutions nationwide that all incorporate this same approach to messaging around engineering that focuses on the contributions of engineers and the role that engineers play in the health, happiness, and safety of society (Haas et al., 2016). Furthermore, undergraduate degree completion in STEM disciplines is overwhelmingly male and white, with white male undergraduates representing 63.6% of conferred STEM degrees in 2018 (National Center for Educational Statistics, 2018). With aims and messaging uniquely situated to attract underrepresented groups in STEM fields, EA programs tend to be more female and more ethnically diverse than the general student population in their respective engineering degree program (Haas et al., 2016). Oftentimes, these EAs strongly believe in the messages that relate to broadening the participation of others in engineering; qualitative, open-ended responses at the end of the AQ often include sentiment that indicates that female students and students of color wish to recruit more young people who also belong to underrepresented groups into engineering (Garner et al., 2018).

Measures of Engineering Students' Perceptions Reviewed for the AQ

Broadly speaking, the qualities of engineering students is a well-established area of educational evaluation in domains such as beliefs, motivations, and self-identifications, though these qualities are rarely measured together with one tool. Of greatest interest to researchers with respect to engineering education has often been areas such as classroom climate, self-efficacy, retention, and perceptions of engineering. As engineering has not attracted and retained significant percentages of historically underrepresented groups, it should not come as a surprise that a number of developed and validated instruments in engineering education are concerned

with these issues (Metz et al., 1999; Marra & Bogue, 2006; Marra et al., 2008; Marra et al., 2009). This section includes the measures that were reviewed and referenced for the purpose of creating the AQ.

The development of the AQ involved looking into a number of these sources with validated measures. The WEPAN climate survey is an instrument designed to tap into perceptions of the classroom climate in engineering (Metz et al., 1999) which is further connected to the Project to Assess Climate in Engineering (PACE). The PACE survey was designed to also measure climate in engineering education and drew upon previous work with the WEPAN project; the validated PACE survey had an average internal consistency reliability statistic of $\alpha = .77$ (Litzer et al., 2014).

Marra and colleagues developed a number of instruments for the self-efficacy and retention of engineering students under the Assessing Women and men in Engineering (AWE) project (2006, 2008, 2009) which has available instruments for research use stored online. The instruments developed for the AWE project, such as the Longitudinal Assessment of Engineering Self-Efficacy (LAESE) have established evidence for validity and reliability, with internal consistency reliability of the LAESE ranging from $\alpha = .72$ to $\alpha = .87$ (Marra & Bogue, 2006).

Another important collection of measures used in the development of the AQ were the messaging surveys developed for the *Changing the Conversation* report, which elicit perceptions from the respondents of the field of engineering (National Academy of Engineering, 2008). The messages that were repurposed for the AQ into survey questions were developed with rigorous qualitative methods to develop the messages and tested on a large-scale quantitative survey (National Academy of Engineering, 2008). The phases of the qualitative message development process involved individual interviews, focus groups, and “youth triads”, which are smaller

concentrated focus groups with adolescents (National Academy of Engineering, 2008). Individual interviews consisted of 12 educators, engineers, and subject area experts; 12 total focus groups were used as a strategy with children, adolescents, and parents (National Academy of Engineering, 2008). The quantitative phase of the *Changing the Conversation* project involved asking for ratings of the messages and taglines through an online survey which was completed by 1,234 individuals (National Academy of Engineering, 2008). The rigorous development process of the *Changing the Conversation* has been provided as the report does not present any specific, quantitative estimates of reliability and validity; however, the methods themselves could be considered high-quality under typical measure development standards, such as using an inductive item generation process and pretesting the targeted sample with a qualitative process (Boateng et al., 2018).

Empirical Measurement in EA Programs

Prior to the development of the AQ, the majority of strategies used by EA programs to evaluate participating students were either qualitative in nature or conducted with unvalidated tools (Garner et al., 2019). While this data was empirical, it was necessary to develop and pilot an omnibus tool with which chapter advisors could systematically assess their students, and the EAN as a whole could compare data and prepare adequately for national trainings (Garner et al., 2019). The development of the AQ included and retained some items from these previously unvalidated measurement tools (Kuhn et al., 2019; Kuhn et al., 2022) allowing these measures to be incorporated in an omnibus instrument with evidence for validity and reliability.

The Current Status of the AQ

Previously, the AQ was developed for the purpose of use with EA programs and has been evaluated for its psychometric properties such as reliability and validity with methods and

techniques such as exploratory factor analysis (Kuhn et al., 2019) and an additional exploratory and confirmatory factor analysis using a random split of the dataset (Kuhn, et al. 2022). However, the final validation of the AQ presented in Kuhn et al. (2022) was not able to retain all items potentially sensitive to differences in gender and ethnicity at the factor analysis stage, with an interim exploration of the measure's psychometric properties and a multiple analysis of variance further suggesting that there were statistically significant racial and ethnic group differences on some subscale composite scores (Kuhn, unpublished pilot study). With items sensitive to potential differences in perceptions of equity unable to satisfy the requirements of consistency between the exploratory and confirmatory factor analyses, a chi square with demographic item responses was used to investigate potential differences on subscales and items across the two randomly sampled groups. The chi square procedure revealed that, aside for theoretical rationale behind potential issues of controlling for gender and ethnicity representation in the groups, a student's year in their postsecondary education and years as an ambassador may have also contributed to differences in variance between the two randomly sampled groups. Therefore, strategies to control for these demographic differences will be explored to verify if items sensitive to demographic differences can be retained in the AQ.

A method of reducing chance for error and increasing the precision of the analysis of the AQ's factor structure is proportionate stratification, which allows different strata of mutually exclusive demographic groups to be sampled proportionately (Salkind, 2010). In this study, proportionate stratification (Salkind, 2010) was employed to ensure that the representation of certain demographics within the AQ are evenly representative across test groups for a revisited exploratory and confirmatory factor analysis. This represents a novel application of this strategy

since there is limited prior research that implements this technique when investigating measure validation.

Summary

Validity of the interpretations of an instrument is not finite; it is something that must constantly be evaluated, and it further behooves researchers to investigate and transparently report standard psychometric evidence of a purportedly validated instrument when using it (Flake et al., 2017). Validity is almost analogous to reduction of error, in many regards; it cannot be perfected, and cannot be absolute. Therefore, the means by which researchers can reduce error in quantitative research is continually investigated and improved upon. If error reduction is a persistent process, perhaps validity should also be considered something in constant development.

In the case of the AQ, there are still possible questions and space for improvement of the instrument, namely the open question of the measure's sensitivity to students' perceptions of differences in equity (Kuhn et al., 2022) as well as reaffirming that it meets the needs of advisors and poses little foreseeable harm to students in the form of adverse consequences of use. To say that the instrument in its present validated form should never again be examined would be misguided and not grounded in the reality of psychometrics. Any future studies that use the AQ in its present form would be encouraged to follow the guidance of Flake and colleagues in carefully examining, justifying, and reporting evidence for the instrument (Flake et al., 2017; Flake & Fried, 2020). Considering the aforementioned diverse nature of EA programs, the importance of equity and broadening participation to the EAN, the focus on role identity of EAs as the foundation upon which the AQ was built, and the core messages in *Changing the Conversation* that are designed to appeal to underrepresented groups in engineering, there is

cause to reexamine the AQ's potential as a tool sensitive to equity differences as well. The body of literature surrounding EAs suggests that diversity and equity are an important part of how EAs see their role. EAs can internalize and strongly identify with the *Changing the Conversation* messages that align with the objectives of the EAN (Garner et al., 2018). Many participants in Garner and colleagues' study also stated clear goals and aspirations that aligned with *Changing the Conversation* or indicated a desire to join an EA program to help broaden participation in engineering (2018).

Finally, without items identified as potentially sensitive to equity and diversity issues in engineering (Kuhn et al., 2019) making it through dimension reduction procedures and the current validation of the AQ leaving diversity as an open question (Kuhn et al., 2022), revisiting the existing data with a stratified sampling by gender, race, and years as an ambassador could either present evidence that an equity subscale is needed in future research with the AQ, or suggest that the AQ in its present validated form as an omnibus measure of EA qualities is satisfactory. Based on arguments for inclusion of consequential validity in the evaluation of measures, identifying potential consequences of use through interview of end users of the AQ was another major goal of the present study.

CHAPTER III

METHODS

Participants

Interview Participants for the Qualitative Research Questions

For the qualitative portion of this study, ten interview participants were purposefully sampled from EA chapter advisors who participated in the initial expert panel review phase of measure development (Kuhn et al, 2019) as well as EA chapter advisors who were not involved in its early stages. The targeted number of interview participants was at least three from the instrument development process and at least five who were previously unaffiliated with the instrument development process. The final sample included four participants who had previous involvement with the development of the AQ and six who were unfamiliar with it. Purposeful sampling was conducted as some EA chapter advisors were involved in previous stages of the instrument's development, but it was also valuable to receive input from chapter advisors who have no prior knowledge of the AQ. Participation was solicited via email with a contact network through the EAN administrator. The advisors contacted were provided with details about the nature of the study and the amount of time that was asked of them along with an electronic informed consent: advisors who completed this informed consent were contacted again to schedule a one-hour Zoom interview. The participants included nine women and one man from both public and private universities across the United States. Two participants were people of color; three participants came from an engineering educational background.

AQ Respondents for the Quantitative Research Question

The quantitative analyses were conducted using responses from 350 undergraduate students in engineering ambassador (EA) programs from across the United States. The demographics of this data set, which was gathered in three rounds between Fall 2018 and Fall

2021, are presented below in Tables 1 - 3. Data was collected at EAN workshops and trainings from current ambassadors the year of administration. The gender distribution of EAs is typical of programs' intentional recruiting of students from underrepresented groups (Hatzell et al., 2013).

Table 1. ***Gender***

	N	**Sample percent**
Male	128	36.6%
Female	222	63.4%

Table 2. ***Race/Ethnicity***

	N	**Sample percent**
Asian	46	13.0%
African American	20	5.7%
Latinx	35	10.0%
Multiracial	23	6.6%
White	221	63.1%

Table 3. ***Years as an ambassador***

	N	**Sample percent**
First year	239	68.3%
Second year	68	19.4%
Third year	33	9.4%
Fourth year or more	10	2.9%

Design

The design of this study was an embedded mixed methods design (Creswell & Plano Clark, 2007) that examined the traditional, predominantly quantitative methods instrument development and validation process (Boateng et. al., 2018) and expanded upon the end of the process with additional qualitative research method via end-user interviews structured to elicit evidence of construct and consequential validity. Qualitative methods and data were used to enhance and improve the current predominantly quantitative process of instrument validation, making this a quantitative approach with an embedded qualitative strand (Creswell & Plano Clark, 2007). Statistical and theoretical evidence demonstrated in an earlier version of the AQ argued for the inclusion of two items sensitive to gender differences (Kuhn et. al., 2019) that the validated AQ did not retain with additional data (Kuhn et al., 2022). Therefore, the quantitative investigation used existing data collected between 2018 and 2021 with an approach that provided proportionate distribution of race, gender, and years as an ambassador between the randomized splits for the EFA and CFA procedures. Meanwhile, semi-structured interviews with an embedded think-aloud approach (Padilla & Leighton, 2017) were used with EA chapter advisors who both took part in the initial panel review and those who did not to determine if there is agreement on the construct validity of the instrument in its present, validated form.

Qualitative Approach

To answer the qualitative research questions of this study, individual interviews were conducted with chapter advisors. A semi-structured interview protocol grounded in Messick's contemporary validity theory (1994) focused on soliciting consequences of use and the extent to which the instrument adequately represented all the constructs that EA chapter advisors want or need to have to be able to support and assess their EAs. These semi-structured interview

questions were developed with a test blueprint to incorporate concepts of Messick's contemporary validity theory (1994), concentrating largely on consequences of use, but also incorporating concepts of concurrent and predictive validity or general construct validity Additional questions were asked of perceptions of the two items that were removed from Kuhn et al.'s (2022) version of the measure.

The first phase of the think-aloud interviews was conducted by showing advisors paragraphs of the AQ's manual over zoom and instructing advisors to share any pertinent thoughts or questions that they had while reading. Prompts avoided drawing attention to any specific section or component of the manual, and instead periodically reminded the participant that sharing any thoughts or questions was encouraged. The second phase of the interview was conducted with references to the test blueprint with interview questions and occasional probes or clarifications of questions. The final phase of the interview with the two omitted questions first allowed advisors to read and review the items and share their initial thoughts, then asked additional questions that arose from these initial responses.

Quantitative Approach

In Kuhn et al.'s (2022) validation study, 41 items were subjected to exploratory and confirmatory factor analyses and two potentially problematic items were removed. Response patterns for these two items that had previously been retained (Kuhn et al., 2019) suggested sensitivity to gender equity differences; therefore, in the present analysis, a proportionate stratification approach with a larger sample re-examined the retention or removal of these items. Prior interim study with the AQ's data identified gender and race as having sensitivities to ECM19 and ECM20, the items that probe for perceptions of equity in engineering; additional interim study identified ceiling effects amongst first-year ambassadors that could skew findings

if overrepresented in a randomly assigned group (Kuhn, unpublished pilot study). Therefore, these three demographic qualities were the central focus of proportionate stratification to ensure variance caused by these respondents' response patterns were not influencing prior findings.

To proportionately stratify the AQ's existing data, all participants were first divided by gender into separate data files. Once in separate files, iterative random number assignment with SPSS's "RV.UNIFORM" function was immediately cross-checked for distribution of race between the two categories randomly assigned with crosstabulations and chi square tests until random assignment by race was nonsignificant and as close to equal distribution as possible. Nonsignificant findings of distribution by race were then examined for the significance of distribution by years as ambassador. Once each gender-separate file was confirmed to have proportionate groupings with no significant differences on all three demographics, all participants were merged into one file once more. The two samples created from this process were tested with iterative rounds of factor analyses and reexamined for response pattern differences with ECM19 and ECM20 using independent samples t-tests.

Measures

Interview protocol. The novel instrument developed for the purpose of this dissertation was an interview protocol grounded in Messick's contemporary validity theory (1994). Questions were developed that solicit consideration for validity, especially the correspondence to the purpose of the instrument and the potential consequences of use of the instrument. A copy of the protocol is included in appendix B. The interview protocol was piloted with a convenience sample of two individuals working in unrelated higher education settings. Interviews were coded using the interview blueprint (see appendix B) as a base code structure for categorizing codes.

This framework included several manifestations of consequences of use and construct validity, as well as areas for general and specific feedback.

AQ survey. Previously, the AQ demonstrated an entire scale Cronbach's Alpha of .94 and subscales that ranged from .72 to .90 (Kuhn et al., 2022). It includes five subscales: Academic Confidence, Engineering Beliefs, Professional Skills, Social Supports, and Ambassadorship. The data used from the AQ was its pilot data, which was collected between 2018 and 2021 (Kuhn et al., 2022). The AQ includes five factors that previously were found to account for 53% of the variance; the five factors identified included Academic Confidence, Engineering Beliefs, Professional Skills, Social Supports, and Ambassadorship (Kuhn et al., 2022). The AQ's validated subscales presented acceptable levels of internal consistency reliability with Cronbach's alpha coefficients ranging from .75 to .90, and an entire scale alpha of .93 (Kuhn et al., 2022). The 41-item scale includes a 5-point Likert-type agreement scale and three additional open-ended items that are used at different points of the academic year, and can be found at the end of the manual in appendix A.

Materials

A brief manual was developed for the purposes of this study to review with advisors during the interview. This manual included a summary of the AQ's components, background, and guidance for interpretation, as well as the scale itself (Kuhn et al., unpublished). The full draft of the brief manual used in advisor interviews can be found in appendix A.

Procedure

Appropriateness

For the qualitative components of the procedure, steps to check for potential researcher bias were taken. This included an external review and testing of the proposed interview protocol

by two higher education professionals, approval from the internal review board prior to conducting research, and auditing of all codes and themes by an external reviewer. The quantitative study involved standard statistical techniques for dimension reduction and detection of subgroup differences.

The Zoom recordings with chapter advisors were deleted after being fully transcribed, checked for accuracy of transcription, and referenced by the external auditor. The universities that the chapter advisors represent are referred to in institution type, size, and geographic region only (e.g., “a large, public northeastern university” or “a small, private college”). For the AQ, the data was de-identified and stored on a secure, private computer. EAs were asked to generate an anonymous unique identifier that followed a replicable pattern so that their responses to the AQ could be tracked from year to year. This generated unique identifier should not disclose any personal information about the EAs.

Qualitative Analysis

Ten EA chapter advisors were interviewed using a semi-structured interview protocol as well as the AQ and its user manual (Appendix A, B). These semi-structured interviews were conducted by the researcher via Zoom and recorded. Closed captioning transcripts were extracted from the video and checked for accuracy against the original recordings. This interview procedure followed a think-aloud methodology, which has been used to gather validity evidence for instruments and examine response patterns of those who take a survey or test (Padilla & Leighton, 2017). However, instead of taking the survey for themselves, participants reviewed a brief draft manual designed for practitioners to understand the use and applications of the AQ, as well as the content of the instrument itself. Following the think-aloud interview procedures outlined by Padilla and Leighton, separate concurrent and retrospective probes (2017) were

developed to solicit responses while chapter advisors review and analyze the draft manual and after the manual review has been completed. The researcher ensured that any concurrent probes were not disruptive to the process of reviewing the manual, nor are led towards any assumptions (Padilla & Leighton, 2017).

The EA chapter advisors were sent a copy of the manual and instrument when being sent a day-before reminder email about the scheduled interview and were asked to briefly skim the manual and instrument. During the interview, the researcher presented paragraphs from the manual one at a time using the screen share function. With the think-aloud portion, any specific/direct probes that may influence or direct the respondent in any specific direction needed to be avoided (Padilla & Leighton, 2017). Therefore, as the advisor progressed through the document, they were occasionally reminded to share aloud their thoughts and feelings about the manual and scale without specific probes that pointed out any portion of the manual. Some advisors had questions in the middle of the first portion of the interview; these were answered when necessary to continue reading, but otherwise written down and answered after the manual review was completed. After a review of the manual and instrument was complete, a separate interview protocol followed with more structured and guided questions to solicit reflections on the consequences of use of the instrument and potential validity of the constructs represented. Lastly, the two items sensitive to gender equity differences and information about these items were presented for discussion, and unstructured questions were generated based on the advisor's initial response to solicit feedback and potential items to include along with the two items in future testing.

After interviews were conducted, they were transcribed and analyzed in with a coding scheme derived from the interview protocol's blueprint for occurrence of codes that established

consensus and uniqueness pertaining to themes of consequences of use, construct validity, and feedback on different portions of the AQ. An external auditor reviewed the codebook derived from the interview protocol, quotes taken from interview transcripts, and theme summaries developed by the researcher and provided feedback on whether conclusions were empirically sound.

Quantitative Analysis

To accomplish the proportionate stratification used in this study, the 350-participant AQ pilot data set was run through iterative processes of random number assignment and chi-square tests for distribution of participants by the demographics of race, gender, and years as an ambassador between the randomly assigned groups. This was first achieved by focusing on gender whereby the entire data set was split into separate files based on male and female responses. Within these two separate files, random number assignment first ensured groups were as close to evenly distributed as possible within both the male file and female files, which was then followed by chi square tests to determine if distribution by race was nonsignificant, and actual values were observed to ensure they were proportionate. Finally, years as ambassador was evaluated with a chi square test to also determine if differences by random group assignment was also nonsignificant. When this was achieved with both the male and female data set, all data was merged into one file once more. This provided exactly proportionate distribution by gender, and distribution by race and years as ambassador. There were no statistically significant differences between the randomized groups in any key demographics the stratification process controlled for (race, gender, years as ambassador).

Afterwards, these data sets were tested with the method of exploratory factor analysis used in both Kuhn et. al. (2019) and Kuhn et. al. (2022) to determine if the earlier of these two

factor structures could be replicated with the proportionately stratified data set. The EFA approach deemed most appropriate for the AQ has been a principal axis factoring extraction with varimax rotation (Field, 2013; Kuhn et al., 2022). As the subscales of the AQ encompass facets of the role identity of EAs, an orthogonal rotation was deemed most appropriate to match its underlying theory (Kuhn et al., 2019).

CHAPTER IV

RESULTS

The following chapter includes the results of the two qualitative research questions and the quantitative research question guiding this dissertation study. Themes will be presented along with sample quotes from interviews to support qualitative findings. Quantitative findings will highlight the outcomes of the stratification process and draw conclusions on the psychometric viability of items sensitive to gender differences.

The first qualitative research question was as follows: “In what ways does the AQ have practical, consequential use for chapter advisors?” Though ten qualitative interviews with EA program chapter advisors, several practical and consequential uses were identified as areas to explore and address in additional modifications and revisions to the AQ’s manual. The second qualitative research question, “What recommendations do EA program advisors make with regards to improving the AQ?”, was also addressed in the framework. The second research question’s categories included feedback on the instrument (both positive and negative) and questions pertaining to the gender equity items, which were included in the third phase of the interview. An external auditor evaluated the codes and themes and their alignment with the quotes identified from the interviews and was in agreement with the conclusions drawn.

The quantitative research question, “How can the construct validity of the AQ be enhanced through proportionate stratification of participants into separate factor analysis groups?”, informed the viability of the items sensitive to equity differences, considering past investigation of the AQ revealed statistically significant differences between groups split randomly for the EFA and CFA procedures and the potential of these items in detecting differences between respondent subgroups.

Qualitative RQ1: In what ways does the AQ have practical, consequential use for chapter advisors?

Areas of consideration for consequences of use of the AQ were identified through the interview process and were compiled into larger themes. Themes included data applications, influence on students, programmatic implications, and data sharing. Themes were also identified that provide further evidence for the construct validity of the AQ, namely complementing existing measures, alignment with areas of interest for measurement, and alignment with existing training procedures for EAs.

Considerations for Consequential Validity: Data Applications.

Data applications defined a group of themes related to the instrument's use and alignment with needs or existing tools presented by advisors. Table 4 below includes an overview of these themes and sample quotes that informed each.

Table 4. ***Sample Quotes for Themes for Data Applications***

Form of Data Applications	**Sample Quote**
Use as pre-post tool to evaluate incoming ambassadors	"Oh, I would love to use the ambassador questionnaire here with my students. The questionnaire would really help me identify what learning outcomes are already kind of inherent in just going. Through the training and in doing the presentations. And then, identify what -- what domains, or what statements kind of need bolstering like if we're going to have expect or want students to succeed in all of these statement areas, then there are going to be some things that are on my end I'll have to modify." "I could see using it to sort of a before and after evaluation to present it to a measure of would just be new ambassadors at the start of the year or new and continuing Ambassador but presented at the start of the year. And the. Maybe mid year but then also, At the end of the year and see how their perceptions of themselves, or of engineering have changed."

Form of Data Applications	Sample Quote
	"Yeah, so I would probably use it as an intake-outtake, because that's where my mind went when I when I heard about the questionnaire. I think my mind went to kind of intake-outtakes type of questionnaire, so I think that that would be my first inclination doesn't mean it's necessarily the best way to use it, but that was definitely where my mind went first."
Frequency of use for pre-post purposes	"I think that it would be interesting to use this when they about halfway through their first year in the program. And then last possibly again like as they enter their last year in the program, or at the end of their last year in the program."
	"Mid-year and then at the end of the year I'd like to see that they have developed that confidence and knowledge about being able to do it. parts of it."
Alignment with existing tools	"We've been doing officer evaluations for several years now about four years ago and one of those evaluations, what it means with the officers, they said, those would be it would have been really helpful to have some of this in the middle of the year, so I had a chance to make changes and could respond to some of these things."
	"We don't usually hire --- we hire freshmen in the spring, then they become sophomores. I have had just I'm just in the recent years, many change majors, so maybe that would give me an indication of how comfortable they are in their major and, well actually none of it -- we don't do that, but we could actually give it in the interview."
Filling need for assessment tools	"I think it's a much more formalized assessment of the program, I think, thus far, we have relied very much on anecdotal evidence of the success of the program whether that's just you know anecdotal. The anecdotal kind of feedback or input from Members, you know, or whether it is."
	"Um, so I don't have something like this, nor do I have time to develop something like this so it would be a nice tool to have. You know, I was just talking about that, like, if you're if you're the advisor and you have to run the program you don't necessarily have time to come up with new training modules or, or, you know, survey and data and evaluation pieces."

Every advisor shared a belief that the AQ could help them identify deficits or areas of need with a new cohort of EAs, and later evaluate if that deficit had been addressed with an additional administration. The AQ as an exact pre-to-post growth evaluation tool was not expressly described within the manual as the survey's capacity to detect long-term change is still under evaluation; regardless, advisors ostensibly viewed it as the survey's purpose. Use as a need evaluation was explicitly described within the manual: "*This survey may be used with new ambassadors and returning ambassadors to assess potential areas of need, such as further professional development workshops or academic and interpersonal support"* (pg. 5). Therefore, this assumption from advisors is predominantly aligned with the instrument's purpose, provided it is used to evaluate needs of current students in ambassador programs.

How often the advisors reported they would ideally administer the AQ varied; some thought to administer it annually, while others described it as best used only upon entry and exit from the outreach program, which could span several years depending on the student. In cases where advisors wished to utilize it more often with each student, the sentiment was related to evaluating the effectiveness of the advisor's practices throughout the year, namely regarding adequate training in aspects of the survey such as an EA's professional skills.

While the process of recruiting, training, and even the form of outreach implemented often varied from program-to-program, the five subscales included in the AQ – academic confidence, engineering beliefs, professional skills, social supports, and ambassadorship – were widely agreed upon as valuable for study with newly recruited EAs. Two advisors also noted that the AQ as a tool could make up for what was largely observational or anecdotal evidence with past cohorts of new students, meaning quantitative data would be available for purposes such as program advocacy and improvement. Four other advisors with access to existing tools used to

evaluate prospective or recently recruited ambassadors consistently identified that the contents of the AQ would align well with these tools; these existing tools often included approaches such as "intake" interviews and faculty EA nomination guidelines/questionnaires. This concept described by advisors is also a question of validity, but concurrent or predictive instead of consequential. The largest area of need advisors agreed upon was the presence of more student data; having the data on EAs' scores on the AQ's subscales was considered inherently valuable by advisors.

Considerations for Consequential Validity: Influence on Students.

Influence on students refers to the potential for use of the AQ to impact the students taking the AQ, positive or negative; the codes comprising this theme and example quotes can be found in Table 5 below.

Table 5. ***Sample Quotes for Themes Related to Influence on Students***

Form of influence on students	**Sample Quote**
Opportunities for student self-reflection	"I think it would be very reflective. I think it would be great to have a baseline of when the students come, and you know, then, especially if you would have you know for a year later, have them come back and respond to it, and then they would. You know, be able, I think that they would learn some things about themselves or about the how the program affected themselves." "It could kind of prompt some self-reflection in them, you know, which I think is advantageous for them. I think it prompts them to do some self-reflection on some of these areas as they're prompted to answer these questions, and you know as they're prompted to respond to the open-ended questions."
Possibility of use as a screening tool	"I mean, we wouldn't - we wouldn't use it as a, "we're going to hire you or not hire you" tool. So there wouldn't be any kind of negative outcome in that way. I don't see any negative to, other than if they take it and they get really down on themselves and start answering some of the questions and feel like maybe they aren't cut out for engineering, but I think that I don't think the survey would be the reason they start thinking that." "I also wonder about it as a screening tool, but sort of as part of the selection process, but given the responsibilities that our ambassadors have high scores on these. don't know that I want to use those in lieu of. or. I think it to me… The

Form of influence on students	Sample Quote
	interview is key for demonstrating that you are someone that I would want to put in front of a prospective student."
Reinforcing self-beliefs and self-perceptions about engineering / perceived low likelihood of adverse impacts	"So I think that if anything it might reinforce that engineers are really important and we do make a difference and this is the best career path that exists and everybody should be an engineer. which, if they if they leave feeling good about their appointment or their by their major and everything that they've done in their time here, they're going to be stronger alumni. which the deans and everybody are going to want because then ultimately donors. I think that would be the kind of the case, it would make for is just kind of reinforcing that ego."
	"I don't feel it would impact them from an emotional standpoint, by any means, maybe in some ways, rewarding and reaffirming because the questions are like you know --- I am confident I can do this you know I have increased my presentation. Oh, and they want my feedback, you know, so here's an opportunity for me to be honest and give them some tips. So maybe reaffirming and seeing feeling seen."
Alternative uses that may impact non-EA students	"I also think about giving in some way to the student body as a whole. Then, I think that Ambassador compared to their classmates. In the same year, and the same major. Would that the ambassadors would have higher ratings on confidence to complete completed agree confident that engineering is right major, and by being selected to the interview process that they are more confident about what they're doing, and its relevance to them. I'm not sure why I find that an interesting question, I'm not sure how I'd use that though."

The most common response to the question of how taking the survey could influence students was conceptualizing the AQ for student self-reflection. These opportunities for reflection included both during the act of taking the survey and having students look at their responses retrospectively to see how they may have changed. Interviewees were not pressed/prompted on how to address maintaining student anonymity with the latter, but the AQ manual does provide guidance on maintaining anonymity of student responses in such a way that responses can still be matched over time. Of the advisors who mentioned sharing students' data with them, one recalled the portion of the manual that pertained to anonymity of responses and

thought further on how to weigh this concern with wanting students to be able to reflect on their own growth. These advisors were not sure how to reconcile this objective of long-term student self-reflection with maintaining anonymity, highlighting that forgoing anonymity of responses could be a potential consequence for students in programs where advisors prioritize this goal.

No advisor identified specific risks for the students that may arise in the act of taking the survey. However, several advisors required an example of what a risk of taking the AQ could look like; the interviewer provided brief examples such as social desirability bias, and that in turn influencing student self-perceptions and self-expectations during survey response. Advisors were then allowed to extrapolate on the possibility of such a scenario; those who asked for clarification in this manner did not think that students' self-perceptions were likely to be altered by the contents of the survey. When potential consequences for students from the act of taking the survey itself were identified, they were largely positive, such as the prosocial engineering beliefs subscale reaffirming positive values about the specific work students do in outreach.

One major concern as an adverse consequence for students from the onset of this study was the potential for misuse as a program entrance screening tool, which could raise questions of construct-irrelevant variance from improper use and possess tangible, negative consequences for students. Two advisors did identify that the AQ could theoretically be used to evaluate which students to admit to their programs, but both subsequently argued against such a use while elaborating further on the idea. Regardless, the fact that any advisors thought to use the AQ in such a way highlights that it *is* a potential outcome and should be addressed in revisions of the manual. Another idea for use of the AQ outside of the intended design with fewer adverse consequences for the students taking the survey was identified by one advisor, which was comparing a sample of non-EA engineering students to EAs using the AQ to see if the AQ can

distinguish between engineering students who participate in outreach versus those who do not. While an interesting concept that could help establish discriminant validity, EA-specific questions are present on the survey that would be incongruent for this purpose.

Considerations for Consequential Validity: Programmatic Implications.

The theme of Programmatic Implications centers on specific, concrete actions that advisors elaborated upon while thinking on how to best inform decisions and plan for the year ahead with their students. Sample quotes of this theme are included in Table 6.

Table 6. ***Sample Quotes for Themes of Programmatic Implications***

Form of Programmatic Implication	**Sample Quote**
Monitor and track student growth	"I would have loved to have had this from the start, to be able to say: these are all the things when you're kind of planning what a year of engineering ambassadors looks like, here's the outcomes that ideally all the students get to have." "And I want to administer it like, oh you're new, tell me where you stand now, and then okay you've done it, one semester. Where you at, you've done it a year where you at. You've left and come back, where are you at. I'm, I really like to have continuation of data, and see okay this student really felt impacted and felt they grew, or the student already felt like they had really strong skills, and then realize they did."
Modify or adjust planned trainings and workshops	"I think that would help us generate ideas around training modules. I think it would - could help with corporate partnership recruitment. Um, I think it would help the network also identify what's most important or valuable to the students, and how to connect that with their trainings." "I mean, I'd be interested in the results in general but some of their professional development questions. I would take those and turn them into professional development opportunities for them. So what came to my mind right now I'm trying to plan this spring conference for the network and I'm thinking, oh, maybe I can go back and look at

Form of Programmatic Implication	Sample Quote
	some of these questions and see, you know, where we might want to focus."

A unanimous conclusion on how to best use the AQ among advisors was some variety of "*pre-post*" usage. This was described in several different ways, such as an "intake-outtake" survey, tracking cohorts across several years, or conducting ongoing monitoring. However, while almost every advisor indicated a desire to document and witness growth, not all had specific means of how the data would inform practices within their program, or other actions and steps they would take using the data. One advisor mentioned actively using the AQ to generate new ideas and areas to cover in training modules within their program that would then be shared at the network level; others mentioned similar sentiments with general trainings, workshops, and seminars. One advisor changed their answer to the question of the utility of the items sensitive to gender equity differences at the end of the interview due to conceptualizing some form of programmatic change that could be addressed based on the findings of the two items. While this consequence seems to be largely positive, it is important to consider that advisors basing plans solely on the AQ could also risk limiting what areas and domains are prioritized for training, or have them discard effective aspects of their program that do not pertain specifically to a subscale of the AQ.

Considerations for Consequential Validity: Data Sharing.

Data sharing as a theme refers to groups and individuals that advisors could foresee sharing the AQ data with, which entails several potential consequences. Sample quotes for this theme are provided in Table 7.

Table 7. ***Sample Quotes for Themes on Data Sharing***

Form of Data Sharing	**Sample Quote**
Faculty sharing	"The process for Ambassador starts with the department, and so they nominate the students in their department, and the students that they nominate are invited to apply and so. So, giving them this as feedback."
	"I mean, with colleagues, I think it would take a lot of colleague trust, because ideally what this would show is the unique value of the engineering Ambassadors Program to improve engineering students abilities to solve problems and have engineer, like sort of like those core engineering beliefs and have better interpersonal communication."
	"So like, we have it on campus. But there's a greater majority of the campus that doesn't know this exists, right - and so I think to get buy in from some of the faculty. Yeah. Having something that is research based to present to them, would resonate. You know, because when I go to hire the students I reach out to the faculty and I say, send this to your students think about who's in your classes recommend it, and some do, because they work with our office and they know."
Administrator sharing	"It means I won't have to come up with an assessment tool. So that's great. It will help me. And again, proving to the department and the college that there's need for the program, the benefits of the program. And then also, because we are a state entity, we have to be able to prove that the money is going. Where we said it's going and the impact that the money is having. So then we could take it to the State Board of Regents and say, See, this is why we spent this money on this because it had this much of an impact."
Corporate partner sharing	"You know as well, it's been really helpful to share this with potential funding sources, I mean not individual data, but you know share kind of overall results with potential funding sources to show the impact of the programs."
	"I never really have anything concrete to share with them, and so I think that that's really valuable even for my current sponsors, even if they're not new sponsors, but my current sponsors, to be able to say, like hey. Look at what we're doing you know your contribution, so this program is supporting all this great work, and so you know I mean engineering companies love to see concrete information."
Network sharing	"And I guess in terms of colleagues I'm thinking about other advisors of other programs. Not just to share maybe what our challenges are and what our successes are in our individual programs, but also potentially being able to kind of aggregate some of that data. You know, as a whole across

Form of Data Sharing	Sample Quote
	our programs look at it like: okay, what impact are our engineering programs having on engineering students."
Research sharing	"In terms of something that I did was work with professors, young professors on their broader impacts on their NSF grants. So it would be something I could show them this program what this program does, and they could put it into their grants. And I think NSF would like that. Fellow advisors, we could compare notes and see how our programs are doing."

Another major elaboration on how data from the AQ on EAs would inform choices and decisions with implications for the program was sharing data with various third parties. One advisor mentioned using it for recruiting new corporate partners to specifically advocate for funding. This advisor indicated that corporate partners tend not to consider how the program is beneficial for the undergraduate students who participate – therefore, they believed sharing the data with partners was valuable for providing evidence of the program's impact. Making others aware of the impact of the program on undergraduate students was also highlighted when advisors responded to the question of sharing data with including supervisory faculty (i.e., deans or university administrators) or peers in different departments across their respective engineering programs. One advisor highlighted that this is especially valuable considering their EA recruitment method, which involves faculty recommendation and nomination. Two advisors also interpreted the question as sharing the data with other advisors for the purpose of collaborating and planning new trainings and workshops. Other faculty members belonging to the advisor's same institution were often noted to be in a position where they were less informed about the value and importance of the EA program. Some advisors believed this data could illustrate how the program is beneficial for students they recommend. Lastly, two advisors mentioned non-

corporate funding, such as their State Board of Regents or larger research grants with new projects pertaining to EAs.

Considerations for Construct Validity: Complementing Existing Tools and Addressing Intended Areas of Measurement.

Forms of construct validity were also identified within interview data, and sample quotes pertaining to concurrent or predictive validity and alignment with intended areas of measurement are included in Table 8.

Table 8. ***Sample Quotes of Forms of Construct Validity with Existing Tools.***

Form of validity	Sample Quote
Predictive for existing end-of-year tools	"Yeah, so I think, like I mentioned we already do an evaluation tool at the end so having one that's been formally vetted it's always nice that we don't have to create anything new. So, I think, I think it would just be a good tool to make sure that you're hiring the right people. And like I said it's already formally created so that's always better than just throwing something together on our end."
Concurrent with existing ongoing tools	"Okay, okay. We do an evaluation here and just thinking about these in relation to that. Evaluation as students -- we have five officers who are the leaders of the program and help take the lead in coordinating activities and planning, different events, and so at the end of each semester. All the members were asked to evaluate those officers and then also the program in general. But we don't that, I have never asked them to evaluate themselves. And that's what it feels like question three is asking about sort of what, maybe that so what they developed the most from but what -- what was meaningful to them and said to. I think an interesting thing to also for me to think about including in our evaluation."

Seven advisors expressed appreciation that the instrument is an available resource and had been piloted and reviewed for evidence of psychometric properties, especially in cases where advisors had not considered including a formal means of measurement prior to reading the AQ Manual. Four advisors were able to follow some of the psychometric properties and expectations

with measurement, while others were less aware of the process of developing a survey; one advisor even expressed that they would not be sure how to analyze the data without outside assistance.

Advisors who already possessed carefully constructed tools for assessing their students in various ways (entry and exit methods) often identified the AQ as a potential means of ascertaining some form of predictive validity for their intake tools, or concurrently validating their own exit interviews, end-of-year surveys, and so on. Despite a consensus on the AQ's alignment with current evaluative needs in EA programs and a sense that currently used tools could align well with the AQ, several advisors pointed out that there are additional areas of measurement that the AQ could address that they don't currently possess means to quantify, such as problem-solving, teamwork, and goal-setting.

Considerations for Construct Validity: Alignment with Student Training.

Three advisors shared specific feedback that indicated that the subscales of the AQ aligned well with things students are explicitly taught. For instance, advisors noted how the Engineering Beliefs questions aligned with the core values of their programs and the sentiment about engineering outreach hopes to spread; this is likely due to the AQ informing the engineering beliefs subscale from the same *Changing the Conversation* report that is used for student training on prosocial messaging. The Social Supports subscale was also noted by one advisor as important for assessing whether the community-building approach used in their program was successful. Professional Skills as a subscale of the AQ was frequently cited as the measure that pertained most to components of ambassador training that was explicitly taught; conversely, Ambassadorship and Academic Confidence were not noted by any advisors as aligning with any specific aspect of EA training. This may be because the two subscales do not

pertain specifically to an area of training but are instead an expected result of participation (Ambassadorship) or an area for monitoring student wellbeing (Academic Confidence).

Qualitative RQ2: What recommendations do EA program advisors make with regards to improving the AQ?

Alongside investigating the AQ's potential consequences of use and construct alignment with EA advisor needs, general feedback was sought from participants on the AQ itself and its draft manual. To align with the third, quantitative research question of this dissertation, feedback on two items with sensitivity to gender equity differences among EAs was reviewed in the third phase of the interview. In this third phase, the interviewer solicited advisors' sense of relevance and utility of the items to their EA program and its mission.

Advisors' General Feedback on the AQ.

All but two advisors indicated in some way that the subscales of the AQ encompassed everything they would wish to measure with their students. By extension, five advisors mentioned they would use the instrument as-is with few or no modifications or adjustments. Despite discussion with some advisors on the value of providing students with their results to reflect upon, most also emphasized that the anonymity of the AQ was a valuable and important component. One advisor further recommended that caution on demographic questions should be present in the manual for especially small programs where certain qualities (i.e., race, gender, years as ambassador) could clearly indicate who is responding. One advisor described that best practices and applications of the manual would be helpful to have available for consideration within the manual. Additionally, two advisors pointed out that "interpersonal skills" may not be the best descriptor for the group of items that loaded onto this subscale; these items seemed to relate more to a sense of belonging and social supports amongst students. This feedback was

integrated into a revision of the instrument for publication purposes separate from this dissertation.

Several other concerns were shared by advisors, but many of these concerns varied greatly depending on the advisor's background and experience with measurement. Two advisors described alternatives to Likert-type scales that could possess greater sensitivity and reduce opportunities for ceiling effects in responses. Two other advisors wanted more opportunities or regular intervals between subscales for written responses and feedback from students, so more detailed qualitative responses could be gathered at the same time. One advisor shared their concerns about the length of the survey, which is over 40 items, and elaborated that this could be time consuming for students and pose a response burden.

Role of the Potential Gender Equity Differences Subscale.

When advisors were asked to evaluate the potential use of items sensitive to gender differences, feelings of advisors on the gender equity subscale were mixed; half of interviewed advisors thought it was valuable to have a scale sensitive to students' perceptions of gender equity, while the other half thought it was unnecessary for the purpose of the AQ. This appears to have divided along educational backgrounds; advisors from an engineering or similar background were more likely to suggest it was unnecessary or irrelevant to how they would use the AQ, while advisors from a humanities or social science background found the items valuable more often (See Table 9). One advisor noted that the responses to these items would also be a personal curiosity. Two advisors changed their minds on the relevance and utility of the scale while discussing it during this portion of the interview; both were initially unsure how it would relate to EA trainings and workshops, but as they talked through the potential applications of the items, both advisors articulated that they could see a way it could play a role in how they design

workshops and be addressed by their program's professional development or teambuilding exercises.

Table 9. ***Advisor Perspective on Equity Items' Utility and Relevance to EA programming.***

Advisor Perspective on Equity Items (approximate summary)	**Postsecondary Educational Background**	**Advisor Gender**
Yes, the equity items have value and could be of use to program purpose	Social science (2), humanities (2), education (1), engineering (1)	6 women
Uncertain/mixed opinion, possibly see value, but not sure how the equity items would relate to program purpose	Engineering (2), education (1)	2 women, 1 man
No, equity items would not be of use for or relate to program purpose	Technical/occupational (1)	1 woman

Amongst the three advisors who were unsure about the use of gender equity specific items in the AQ, two of the advisors who expressed this sentiment connected it in some way to uncertainty on how to impact change on students on these items, citing that gender equity is a universal issue in engineering and is difficult to change ambassadors' perceptions on during regular trainings and workshops. This sentiment was also often relayed in the sense that it would be difficult to see change over time on such a subscale because of the challenges addressing the issue in regular EA workshops. Along with the quantitative findings, this perhaps suggests that the gender equity items would best stand alone as a new measure of sense of gender equity in the field of engineering for students participating in engineering outreach.

In the cases that advisors provided positive or mixed opinions on the utility and applications of the gender equity items, they were also asked if they had suggestions for potential items that could accompany a new gender equity sensitive subscale.

Advisors' Recommendations for Additional Items.

Advisors who indicated being able to see some use in the construct suggested by the two items that were sensitive to gender differences were queried about potential items that could compliment the two existing items. The ideas and language provided were grouped thematically and summarized by the researcher into a brief scale of six items for future piloting, as shown in Table 10.

Table 10. ***Advisor-Suggested Items for Future Piloting with Existing Items***

New Item description	Status
I am confident I will be treated fairly on the job in engineering.	Existing
I am confident I will be given the same opportunities for pay raises and promotions as my fellow workers if I enter engineering.	Existing
I am comfortable sharing my diverse experiences with others in engineering.	New
I am confident I will be treated fairly in my engineering courses.	New
I am confident I will be treated fairly in any internship experiences I have in engineering.	New
I believe I can help make engineering a field that values diverse experiences.	New

Quantitative RQ: How can the construct validity of the AQ be enhanced through proportionate stratification of participants into separate factor analysis groups?

The chi square tests used indicated that with the stratification of participants into the CFA and EFA evaluation groups by gender, race, and years as an ambassador, there were no significant differences between the split groups, as shown in Table 11 below.

Table 11. ***Results of Stratification of Group Assignment***

Demographic Indicator		***N* group 1**	***N* group 2**	χ^2	***p***
Sex				0	1.00
	Male	61	61		
	Female	111	111		
Race				6.96	0.32
	Af. Amer. or Black	13	7		
	Asian	24	22		
	Hispanic/Latino	17	18		
	White	110	111		
	Two or more	11	12		
	Other (write-in)	0	3		
	Not respond	0	2		
Years as				1.87	0.60
ambass.	First year	120	119		
	Second year	36	32		
	Third year	16	17		
	Fourth or more	3	7		

Iterative diagnostic exploratory factor analyses with principal axis factoring and Varimax rotation (Field, 2013) revealed that the piloted items ECM19 (“I am confident I will be treated fairly on the job in engineering”) and ECM20 (“I am confident I will be given the same opportunities for pay raises and promotions as my fellow workers if I enter engineering”) would only load as a two-item factor when a sixth factor was allowed, and remaining items loaded similarly to the original five-factor structure described in Kuhn et. al. (2022) that excluded these two items. ECM19 and ECM20 loaded at values of 0.73 and 0.81 respectively under this condition. When five factors were forced, the resulting factor structure was uninterpretable with ECM19 and ECM20 remaining in the pool of items. As a two-item factor is typically not viable

and factor structures must be interpretable in a meaningful way (Watkins, 2018) the retention of these items with the stratification process was deemed unsuccessful, also reinforcing the factor structure in Kuhn et al. (2022) that excludes them.

Despite the process's inability to retain the items of interest, these items are still found to be sensitive to gender differences in both stratified random samples, as seen in Table 12.

Table 12. ***T-tests of Gender Differences with Items of Interest by Stratified Sample.***

		t	***df***	***p***	***d (effect size)***
Sample 1	ECM19	6.54	148.96	< .001	.98
	ECM20	6.87	143.35	< .001	1.04
Sample 2	ECM19	3.92	140.25	< .001	.60
	ECM20	6.14	155.60	< .001	.90

Along with the mixed opinions of advisors on the utility and relevance of these two items to the AQ, this finding suggests that items ECM19 and ECM20 would be best-suited in a new equity-focused scale in engineering education, or additional sixth subscale for the AQ. In sum, the approach of proportionate stratification to ensure the data sets for the exploratory and confirmatory analyses had as little difference as possible in the demographics of race, gender, and years as an ambassador verified that the validation omitting the two items in Kuhn et al. (2022) should be retained.

CHAPTER V

DISCUSSION

In the process of gathering evidence for the validity of an instrument, one must consider as many sources as possible. This study aimed to collect evidence for consequential validity, which is one of the facets of Messick's (1994) Contemporary Validity Theory with few accepted methods of collection. Facets of consequential validity informed the design of a user interview protocol, and while searching for potential consequences of use of the AQ, the construct validity of two items sensitive to gender differences was also explored. The qualitative feedback ascertained for two items in a theoretical perceived equity-focused scale was contrasted with the quantitative portion of this study, which aimed to test if the two items could be retained if the data sets used for the exploratory and confirmatory analyses were similar in race, gender, and years as an ambassador. Synthesizing the mixed advisor opinions with the outcomes of the proportionate stratification, future research should move forward with tests of an entirely new scale related to students' perceptions of gender equity and diversity in engineering outreach.

Along with a method of data stratification, this dissertation study endeavored to explore a novel method for collecting early evidence of potential consequences of use of an instrument. With potential consequences identified through the interview process, suggestions for modification and expansion of the survey's user manual can be made. Development and implementation of a user manual for survey instruments is not a widespread practice itself; however, wider development and implementation of user manuals for survey instruments could be the ideal way to preempt adverse consequences of use. Referencing themes described several changes and suggestions for adjustments to the manual's language can be made.

Addressing Consequences of Use with Revisions to the AQ Manual

Clear direction needs to be provided and elaborated upon for EA advisors regarding when to use the AQ and what to expect from its outcomes. For example, based on the tendency towards high scores on individual items described in Kuhn et. al. (2022) incoming students could be overconfident, presenting ceiling effects the first time the survey is taken (Kuhn et al., 2019), only to drop to a more realistic and accurate score later; advisors should have this consideration made clear to them in the survey's instructions to not misinterpret scores that are either stable or decline while looking for data on student growth. Along similar lines, clear instructions for analysis procedures and interpretation should be provided within the manual, such as how to derive mean or sum scores for subscales. Advisors should also be made aware that there is no normative data to compare their students' results to; interpretation of the results with their students will likely come in the form of identification of room for improvement instead of searching for deficits. These considerations will most likely address the largest theme amongst advisors, which was use of the AQ to identify areas for growth amongst their students in constructs of interest.

Another consideration for additions or edits to the AQ's manual to address foreseeable consequences of use is the potential for the instrument to influence students and their experiences, despite advisors believing the potential adverse impacts would be limited or largely positive, such as creating opportunities for self-reflection. Though no advisor said it would be a primary use, two did identify that the AQ could theoretically be used as a tool to screen students prior to joining an ambassador program; this is a major divergence from the AQ's designed purpose and would be a form of use with potentially detrimental consequences for students. This form of use should be specifically described as inappropriate within the AQ user manual to

preempt such uses. Following Messick's description of consequential validity, this form of use would be prone to construct-irrelevant variance, and the actions taken would have adverse implications for the students taking the survey (1994, p. 15).

Applications to the advisors' practices should also be considered and addressed within the user manual. In the act of providing a standardized measure of some kind for a group of stakeholders, these stakeholders behaviors may change; EA advisors echoed this phenomenon often seen in achievement testing by describing use of the AQ to make specific modifications and changes to how trainings and workshops are structured. This may mean that the focus of the trainings themselves may shift, or some valuable trainings may be omitted in favor of new ones that reflect constructs found on the AQ. Advisors should be made aware that the constructs addressed by the AQ are not exhaustive, and there are still more skills and issues with students that may require their guidance and support.

Anonymity posed an issue for advisors who wished to also share the results with students over time so they would be able to reflect on their own growth, and protection of this anonymity for students was already a consideration from the potential to share their responses with groups such as other faculty, other programs, university administrators, and funders. Though some direction in the manual is provided for advisors to make an identifier that will help them track students' responses over time, more could be provided so that advisors fully understand why anonymity is valuable and potential alternatives to address the issue of sharing growth with students.

Methodological Implications of Findings

While gathering data for the potential consequences of use of the AQ, several extensions of methods for data collection and analysis were tested. This included the design of an interview

procedure for the purpose of forecasting consequences of use of an instrument and a method of proportionately stratifying a data set for separate factor analyses.

Despite think-aloud procedures providing valuable evidence for construct validity with interviews where the interviewee is answering questions of critical thinking (Padilla & Leighton, 2017) the think-aloud procedure in the first phase of the interview was largely a poor source of relevant data for consequences of use and construct validity, and only occasionally provided general feedback. General feedback from the first phase was often underscored by clarification questions asked while reviewing the manual. Nevertheless, the strategy likely posed as an effective exercise in preparation for the second phase; reviewing the manual with the interviewer was cited often in response to structured questions, and interviewees occasionally asked for it to be screenshared again. The most beneficial part of the interview procedure for ascertaining evidence of construct and consequential validity was the second phase of the interview, wherein respondents answered semi-structured questions informed by Messick's Contemporary Validity Theory (1994). This suggests that probes based on theory provide the most relevant data for exploring potential consequences, including ones the interviewer had not anticipated, such as advisors stating they would consider changing plans for workshops to align them with the AQ.

The second method tested with this dissertation was the process of proportionately stratifying a data set to ensure there was no variance between two randomly split data sets. Past research with the AQ suggested that the instrument itself was potentially sensitive to differences based on demographics, with two items that were not retained in the factor structure presenting as especially sensitive to differences in gender. These past findings highlighted a need for such a method; this process involved iterative randomizations to with repeated chi square tests to ensure the splits were first as even as possible by gender, then as even as possible by race, and finally as

even as possible by years as an ambassador. These proportionately stratified groups did not change that the two items sensitive to gender differences would not load with the rest of the AQ in its present state, thus confirming that the issue with past analyses was the items themselves and not unequal variance in group demographics. While this method only verified past findings, it also highlighted along with the mixed opinions of advisors that a separate scale or subscale would be needed if the field of engineering outreach wishes to investigate ambassadors' sense of inequity in the field.

Limitations

Several limitations are present in this dissertation study. First, it is possible that additional potential consequences were not identified as interview participants did not review survey data to practice drawing conclusions from the survey; this is a potential exercise that could be added in the future to this procedure in instrument manual development. Desirability may have occurred with advisors' responses, and it is hard to say if the interview participants' responses were at all informed by a belief about the instrument's intended and unintended purposes from the designers. Further development in this process in the future could involve greater emphasis on original feedback, especially with participants who possess varying degrees of familiarity with measure development. It is also important to consider that even with the suggested edits and changes to make in the AQ's manual that attempt to address and control for adverse consequences, it is not guaranteed that those who wish to use the AQ will read and follow all guidance provided.

Limitations also exist for the quantitative portion of this study. At present, the quantitative procedure involves iterative rounds of sorting and analysis with chi square tests; more precise, automated procedures need to be developed for this method to reduce the role of

subjective evaluation of an adequate non-significant chi square for each demographic reviewed. This result may also be hard to achieve with larger data sets.

Significance

At present, few studies have used qualitative methods in the process of evaluating the quality of an instrument after validation and during the preparation of a manual designed to guide the instrument's use. Furthermore, widespread development of user manuals for instruments in educational psychology is not yet a standard practice; this study aims to argue for the use of such manuals as a tool for both evaluating and controlling for potential adverse consequences of use. Researchers using and following the procedures and guidance laid out in a well-designed manual could both address many foreseeable issues of consequential validity while also addressing many of the concerns related to high-quality reporting of instrumentation outlined by Flake and Fried (2020). Persistent issues plague educational and psychological measurement in this regard; instruments are often used in studies with little transparency on how they were modified and justification for the study's relation to the instrument's constructs (Flake & Fried, 2020). Practitioners in the field of educational measurement should consider the new method at the core of this study, namely instituting manuals for instrument users and testing said manuals with the demographic of persons who would demonstrate interest in using the tool.

Using this approach, several foreseeable consequences were identified that could also be addressed with additions and changes within the AQ's manual, and clear instructions will ideally encourage appropriate use and interpretation by those who would use it. However, additional consequences could appear in the future that were previously unidentified, but these could be addressed with the cyclical nature of the measurement development and validation process. One consequence in particular was unexpected, and difficult to categorize as positive or negative: the

potential for advisors to change their plans and workshops to meet the anticipated areas of measurement found in the AQ. As measuring certain qualities inherently makes people more aware of said qualities, this consequence was likely to appear regardless of steps taken; it is also difficult to say whether the outcomes of this consequence are positive or negative. Previously, this phenomenon has been identified and termed "washback" (Messick, 1996) in application to achievement-based settings.

This dissertation study also endeavored to compare qualitative and quantitative data to determine if items removed from previous iterations of the AQ found in Kuhn et. al. (2019) that were known to be sensitive to gender differences could be reintegrated. With the use of proportionate stratification of data into randomly split groups, it was determined that the existing factor structure that removed these items and was validated in Kuhn et. al. (2022) was appropriate. Considering the mixed feedback from advisors on the utility of the items in the main survey, these two findings suggest that equity-focused items for use with EAs may be best suited for an additional subscale, or an entirely new scale separate from AQ administration. Additional development with the AQ is possible in other ways: at present, perceived action possibilities from the instrument's theoretical framework are only integrated into the AQ via qualitative, open-ended questions, and a sixth subscale involving perceptions of future equitable treatment could also integrate this concept. Additional development of the AQ will come from feedback from the advisors, which will be used to develop additional items that could be piloted along with the original two items in future research. Future research should also consider how the process of measure development is rarely definitive, and evaluation of instruments to improve their performance and reduce potential harms arising from consequences of use should be an ongoing process.

APPENDIX A

DRAFT BRIEF MANUAL

The Ambassador Questionnaire (AQ): A Tool for Evaluating Students Conducting Outreach

Developed by Melissa G. Kuhn, Joanna K. Garner, and Shanan Chappell Moots

Part 1: What is the Ambassador Questionnaire (AQ)?

The AQ is a 41-item survey with five subscales that was developed for use by both researchers and practitioners who work with groups of students conducting outreach activities in engineering. This survey was developed by reviewing existing surveys made for researchers and practitioners in engineering education and adapting them to a role identity framework. The development of the AQ involved chapter advisors from engineering ambassador programs and educational psychologists prior to its piloting to undergraduate students in engineering ambassador programs in 2018. Pilot data was collected from students between the 2018-2019 and 2019-2020 school years. The pilot data was used to check the performance of the survey's items and subscales and refine the AQ into a final, valid and reliable version for wider use by the researchers and practitioners it was developed for. The validated form of the AQ is currently under peer review for publication.

The AQ is best used with a 5-point, Likert-type agreement scale. Items are anchored at the following points: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree. The five subscales of the AQ include the following: 1) Professional Skills (15 items), 2) Engineering Beliefs (9 items), 3) Academic Confidence (10 items), 4) Interpersonal Skills (5 items), and 5) Ambassadorship (3 items). There are a small number of open-ended items that may also be used to gather qualitative responses from students.

Part 2: What are the Components of the Ambassador Questionnaire (AQ)?

The 42-item survey can be broken down into five subscales with different components relating to the self-perceptions, self-definitions, beliefs, goals, and action possibilities undergraduate students conducting engineering outreach may experience. These five subscales vary in length and were informed by and adapted from existing measures commonly used in engineering education, such as instruments from the Project to Assess Climate in Engineering (PACE) and the Assessing Women and Men in Engineering (AWE) project. The survey is also informed by the *Changing the Conversation* report published by the National Academy of Engineering.

The *Professional skills* subscale consists of 15 items and asks students to consider and evaluate the professional skills they are developing, such as organizing and presenting information to others, communicating effectively, and working collaboratively with other ambassadors. This subscale had an internal consistency reliability of $\alpha = .92$.

The *Engineering Beliefs* subscale consists of 9 items that involve beliefs and assertions about the field of engineering. If used with a broad population of students conducting outreach, modification to the verbiage in this subscale is recommended. Engineering beliefs includes ideas such as engineers are problem solvers, engineers make a difference, and engineers are creative. This subscale had an internal consistency reliability of $\alpha = .87$.

The *Academic Confidence* subscale consists of 10 items that evaluate students' academic self-perceptions and beliefs about academics, including ideas such as confidence in degree completion, feeling satisfied with major choice, and perceived future success and applications of engineering once in the field. This subscale had an internal consistency reliability of $\alpha = .89$.

The *Interpersonal Skills* subscale consists of 5 items and asks students to consider feelings of belonging, relating to peers, seeking mentorship, and making friends. This subscale had an internal consistency reliability of $\alpha = .75$.

The *Ambassadorship* subscale consists of 3 items and ask students to specifically situate how their participation as an ambassador and outreach activities will influence their understanding of engineering, their understanding of their career goals, and their confidence in their ability to succeed in engineering. This subscale had an internal consistency reliability of $\alpha = .88$.

Open-ended response items

Two open-ended response questions are recommended for use in the fall semester of a given school year, and two additional items for use in the spring semester. The items are as follows:

1) What were your main reasons for choosing to participate in the Ambassadors program?

2) What are your plans after graduation? (Indicate the job or further education you plan to pursue)

3) What was your most meaningful experience in your Ambassadors program?

4) What suggestions do you have for improving your Ambassadors program?

Part 3: How do I Use and Interpret the Ambassador Questionnaire (AQ)?

The AQ was designed to be used with digital survey platforms such as Qualtrics or Google Forms. To properly scale the AQ, it is recommended to label all five anchor points with the corresponding agreement verbiage. Students should enter a unique identifier that follows a Self-Generated Identification Code (SGIC) format at the start of the survey so changes in their responses can be compared from fall to spring. The recommended format for this SGIC is 1. First letter of mother's first name, 2. How many older brothers you have, 3. What is the number that represents your birth month, and 4. What is the first letter of your middle name (if none, use X). Collection of demographics is also recommended when reasonable; pilot testing of the AQ asked students to indicate their race/ethnicity, gender, years as an ambassador, and level in college. If working with a small group of ambassadors, collection of only gender and race may be advised to ensure student anonymity. Anonymity is needed to ensure student responses are not subject to responding as they believe their chapter advisor or researchers want them to respond.

Once data has been collected from students, it is recommended that the data be reviewed in a statistical program such as Microsoft Excel or IBM-SPSS that can calculate a minimum of composite scores on subscales and provide descriptive statistics. In part 4, please note that not all items are in order of the subscale they belong to; be sure to correctly group the items when calculating composite scores. For practitioners, subscale score means can provide an easy summary of responses that can be compared from year to year. Other simple analyses can be run to identify specific areas of need by demographic groups, such as comparing responses on subscales or specific items using crosstabs. For researchers, more sophisticated analyses such as multiple analysis of variance (MANOVA) are recommended. It is also recommended that researcher reporting include validity and reliability evidence for a new study that has used the AQ.

This survey may be used with new ambassadors and returning ambassadors to assess potential areas of need, such as further professional development workshops or academic and interpersonal support. It is also usable with future research studies involving postsecondary students conducting outreach activities.

Part 4: The Ambassador Questionnaire

Item	Subscale
1. I am confident I can succeed in my engineering courses.	AC
2. I am confident I can succeed in my math courses.	AC
3. I can complete the math requirements for most engineering majors.	AC
4. I am confident I will complete an engineering degree in my preferred major.	AC
5. I am confident that engineering is the right major for me.	AC
6. I am satisfied with my decision about my specific engineering major.	AC
7. I am confident I will complete an engineering degree.	AC
8. It is my choice to study engineering.	AC
9. A degree in engineering will allow me to get a job where I can use my talents and creativity.	AC
10. Someone like me can succeed in an engineering career.	AC
11. A degree in engineering will allow me to obtain a job that I like.	AC
12. My participation as an ambassador will lead to a better understanding of engineering.	AMB
13. My participation as an ambassador will lead to a better understanding of my own career goals.	AMB
14. My participation as an ambassador will make me more confident about my ability to succeed in engineering.	AMB
15. Engineers are creative.	EB
16. Engineers are problem solvers.	EB
17. Engineering is essential to our health, happiness, and safety.	EB
18. Engineering connects science to the real world.	EB
19. Engineers help to make the world a better place.	EB
20. Engineers help shape the future.	EB
21. Engineers make a world of difference.	EB
22. Engineers are entrepreneurial and innovative.	EB
23. It is important that the field of engineering is diverse.	EB
24. I can weigh alternative perspectives in solving a problem.	PS

Item	Subscale
25. I can develop ways to resolve conflict and reach agreement in a group.	PS
26. If a team I am working on has a problem, I can help the team resolve the issue.	PS
27. I can synthesize multiple points of view that arise during group problem solving.	PS
28. I can organize information so that it is easily understandable to others.	PS
29. I can communicate effectively.	PS
30. I can create engaging presentations.	PS
31. I can deliver engaging outreach presentations.	PS
32. I work cooperatively with other ambassadors on our duties and tasks.	PS
33. I know how to integrate messages from ‘‘Changing the Conversation’’ into an outreach presentation.	PS
34. I am confident in my leadership skills.	PS
35. I can provide a helpful and appropriate critique to a fellow Ambassador’s presentation.	PS
36. I can select an appropriate slide design to suit the audience of my presentation.	PS
37. At this university, I feel like I am part of a group.	SS
38. If I have a problem, there is someone at this university I can talk to.	SS
39. I have a mentor to whom I can turn for advice.	SS
40. I can relate to the people around me in academic settings.	SS
41. It is easy for me to make friends.	SS

APPENDIX B

Interview Protocol

Question	**Notes / Theorized codes**	**Construct / Target**
1. What stood out to you about the way the ambassador questionnaire is described in the manual?	Construct validity of instrument: attempting to determine clarity/coherence of manual	Feedback, clarity or comprehension
2. How can you see yourself using the ambassador questionnaire with your students? In what ways will the questionnaire be useful with your students? How will you use the data?	Consequences of use – data applications Guiding program decision making (workshops, interventions)	Consequences of use
3. In what ways do you think your students would respond to the ambassador questionnaire?	Consequences of use – impact on students Response patterns that may have beneficial or detrimental consequences for the students? *go back to the manual and include the blurb for the beginning of Qualtrics*	Consequences of use
4. How can you see yourself using the ambassador questionnaire with your colleagues?	Consequences of use – utility with faculty/staff colleagues Program value advocacy? Evaluative proof? Soliciting more funding? Network level vs. internal in college?	Consequences of use
5. How can you see the ambassador questionnaire being improved?	General feedback – residual needs? Utility validity?	Feedback, utility
6. How do you think the ambassador questionnaire could meet the needs of your program?	Consequences of use – utility Impact on current areas of deficit	Consequences of use

Question	Notes / Theorized codes	Construct / Target
7. What did you like about the ambassador questionnaire and the manual?	General feedback	Feedback, general

Third phase – introduce and focus on the two more items. Adapt questions based on responses to reading this section.

Omitted from manual:

Outside of the 42 validated scale items and open-ended items, there are two items with some preliminary evidence to suggest they are sensitive to gender differences that practitioners can opt to use, but were not validated for research use with the AQ. These two items were not able to be included as they acted as one sole, two-item factor; it is not considered standard practice to validate a two-item scale.

Optional items

There are two items that were not validated for the AQ that demonstrated statistically significant sensitivity to gender differences using exploratory independent samples t-tests. As these items did not hold for the validated scale, they are not included in the validated version. These items require further development and research to be used with a validated ambassador survey.

1) I am confident I will be treated fairly on the job in engineering.
2) I am confident I will be given the same opportunities for pay raises and promotions as my fellow workers if I enter engineering.

APPENDIX C

IRB Exemption Letter

OLD DOMINION UNIVERSITY

OFFICE OF THE VICE PRESIDENT FOR RESEARCH

aadlac INTERNATIONAL
PROUDLY, AN AAALAC ACCREDITED PROGRAM

Physical Address
4111 Monarch Way, Suite 203
Norfolk, Virginia 23508
Mailing Address
Office of Research
1 Old Dominion University
Norfolk, Virginia 23529
Phone(757) 683-3460
Fax(757) 683-5902

DATE:	August 4, 2021
TO:	Joanna Garner
FROM:	Old Dominion University Education Human Subjects Review Committee
PROJECT TITLE:	[1780243-1] Investigating the process of consequential validity with the ambassador questionnaire
REFERENCE #:	
SUBMISSION TYPE:	New Project
ACTION:	DETERMINATION OF EXEMPT STATUS
DECISION DATE:	
REVIEW CATEGORY:	Exemption category # 2

Thank you for your submission of New Project materials for this project. The Old Dominion University Education Human Subjects Review Committee has determined this project is EXEMPT FROM IRB REVIEW according to federal regulations.

We will retain a copy of this correspondence within our records.

If you have any questions, please contact John Baaki at (757) 683-5491 or jbaaki@odu.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within Old Dominion University Education Human Subjects Review Committee's records.

REFERENCES

Anagnos, T., Lyman-Hold, A., Marin-Artieda, C. & Momsen, E. (2014). Impact of engineering ambassador programs on student development. *Journal of STEM Education: Innovations and Research,* 15(3), pp. 14-20.

Assessing Women and Men in Engineering (2008). STEM assessment tools: Retention surveys: Students leaving engineering. Available: https://www.engr.psu.edu/awe/secured/director/retention/leaving.aspx

Assessing Women and Men in Engineering (2008). STEM assessment tools: LAESE: Longitudinal assessment of engineering self-efficacy. Available: https://www.engr.psu.edu/awe/secured/director/diversity/efficacy.aspx.

Astin, A.W. & Sax, L.J. (1998). How undergraduates are affected by service participation. *Journal of College Student Development, 39*(3), pp. 251-263.

Bates, J.S., Krapcho, K.J. & Orantes, C. (2014, June). How to recruit and retain students using an engineering ambassador program. *Paper presented at the ASEE Annual Conference and Exposition,* Indianapolis, IN.

Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quinonez, H.R., & Young, S.L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health 6,* 1-13.

Canney, N.E. and Bielfefeldt, A.R. (2015). Volunteerism in engineering students and its relation to social responsibility. *Paper presented at the ASEE Annual Conference and Exposition*, Seattle, WA.

Cizek, G.J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Creswell, J.W. & Plano Clark, V.L (2007). *Designing and conducting mixed methods research.* SAGE Publications.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics.* SAGE Publications.

Flake, J.K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social, Psychological, and Personality Science, 1*(9). DOI: 10.1177/1948550617693063

Flake, J.K. & Fried, E.I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Sciences, 3*(4). https://doi.org/10.1177/2515245920952393

Furr, R.M. & Bacharach, V.R. (2014). *Psychometrics: An introduction.* SAGE Publications.

Garner, J.K., Haas, C., Alley, M., & Kaplan, A. (2018). The emergence of outreach ambassador role identities in undergraduate engineering students. *Journal of STEM Outreach, 1*(2) p. 1-18. Dec. 2018.

Garner, J.K., Alley, M. & Kuhn, M.G. (2019, June). Common practices in undergraduate engineering outreach [Paper Presentation]. Presented at *the Annual Meeting and Exposition of the American Society for Engineering Education,* Tampa, FL, United States.

Garner, J.K., Thole, K.A., & Alley, M. (2020, June). Work in progress (WIP): A systematic review describing impacts on engineering undergraduates who participate in outreach [Paper Presentation]. Presented at *the Annual Meeting and Exposition of the American Society for Engineering Education,* Virtual.

Gaskin, C.J., Lambert, S.D., Bowe, S.J. & Orellana, L. (2017). Why sample selection matters in exploratory factor analysis: Implications for the 12-item World Health Organization

Disability Assessment schedule 2.0. *BMC Medical Research Methodology*, *17*(40). https://doi.org/10.1186/s12874-017-0309-5

Godwin, A., Potvin, G., Hazari, Z., & Lock, R. (2016). Identity, critical agency, and engineering: An affective model for predicting engineering as a career choice. *Journal of Engineering Education, 105,* (2), 312-340.

Green, A. (2018). The influence of involvement in a widening participation outreach program on student ambassadors' retention and success. *Student Success*, *9*(3), pp. 25-36.

Habash, R.W. & Suurtamm, C. (2010). Engaging high school and engineering students: A multifaceted outreach program based on a mechatronics platform. *IEEE Transactions on Education, 53*(1), pp. 136-143.

Haas, C., Alley, M. & Garner, J.K. (2016, June). *Engineering Ambassadors Network: Goals, successes and challenges in growing the EAN* [Paper presentation]. American Society for Engineering Education 123rd Annual Conference and Exposition, New Orleans, LA, United States.

Hatzell, J.G., Marshall, M., Alley, M., Thole, K.A., Haas, C., Engel, S.R., & Garner, J.K. (2013, June). *Engineering Ambassadors Network: Professional development programs with an outreach focus* [Paper presentation]. American Society for Engineering Education 120th Annual Conference and Exposition, Atlanta, GA, United States.

Huff, J.L., Zoltowski, C.B., & Oakes, W.C. (2016). Preparing engineers for the workplace through service learning: Perceptions of EPICS alumni. *Journal of Engineering Education, 105*(1), pp. 43-69.

Kaplan, A. & Garner, J.K. (2017). *Dynamic Systems Model of Role Identity (DSMRI) Analysis Guide & Codebook*. Personal Communication.

Kuhn, M.G., Garner, J.K., & Chappell Moots, S. (2019) *Work in progress: Development and validation of the Ambassador Questionnaire* [Paper presentation]. American Society for Engineering Education 126th Annual Conference and Exposition, Tampa FL, United States.

Kuhn, M.G., Garner, J.K., & Chappell Moots, S. (2022). Validation of the ambassadorship questionnaire for undergraduate students conducting engineering outreach. *International Journal of Engineering Education*, 38(5), pp. 1223-1242.

Kuhn, M.G., Garner, J.K., & Chappell Moots, S. (Unpublished). The Ambassador Questionnaire: A tool for evaluating students conducting engineering outreach. Brief Manual.

Kuhn, M.G. (Unpublished). Interim investigation of the psychometric properties of the AQ. Included chi square tests between randomly assigned EFA and CFA conditions and observation of ceiling effects in first year ambassador responses.

Kwiatkowski, S., Klein, K.C., & Mason, N.A. (2019). Impact of a pharmacy student organization on the admissions experience and student professional development. *American Journal of Pharmaceutical Education, 83*(7), pp. 1471-1479.

Laanan, F. (2010). Studying transfer students: Part 1: Instrument design and implications, *Community College Journal of Research and Practice 28* (4). pp. 331-351.

Laursen, S.L., Thiry, H., & Liston, C.S. (2012). The impact of a university-based school science outreach program on graduate student participants' career paths and professional socialization. *Journal of Higher Education Outreach and Engagement, 16*(2) pp. 47-77.

Litchfield, K. & Javernick-Will, A. (2015). "I am an engineer AND": A mixed methods study of socially engaged engineers. *Journal of Engineering Education 104* (4), 393-416.

Litzler, E., Samuelson, C.C., & Lorah, J.A. (2014). Breaking it down: Engineering student STEM confidence at the intersection of race and gender. *Research in Higher Education,* 55(8), pp. 810-832.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Education Research, 11*, pp. 253-388.

Marsh, H.W. & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), pp. 1187-1197.

Marsh, H.W., Abu-Hilal, M.M., Abdelfattah, F., Xu, M.K., Abduljabbar, A.S., Morin, A.J., Leung, K.C., & Nagengast, B. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology, 105*(1), pp. 108-128.

Marra, R.M., & Bogue, B. (2006). Women engineering students' self-efficacy: A longitudinal multi-institution study [Paper Presentation]. *Women in Engineering Programs and Advocates Network (WEPAN) Conference. Pittsburg, PA, USA, June 11-16.*

Marra, R.M., Bogue, B. Shen, D. (2008). Engineering classroom environments: Examining differences by gender and departments. *2008 American Society for Engineering Education Annual Conference & Exposition, Pittsburg, PA, USA, June 22-25.* https://www.engr.psu.edu/awe/misc/ResearchPagePDFs/AWISE_ASEE08_final.pdf

Marra, R.M., Rodgers, K.A., Shen, D., & Bogue, B. (2009). Women engineering students and self-efficacy: A multi-year, multi-institution study of women engineering student self-efficacy. *Journal of Engineering Education,* 98 (1), pp. 27-38.

Messick, S. (1988). Validity. In Linn, R.L. (ed.) *Educational Measurement.*

Messick, S. (1994, June). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Presented at *the Conference on Contemporary Psychological Assessment,* Stockholm, Sweden.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Messick, S. (1996). *Validity and washback in language testing.* Educational Testing Service.

Metz, S., Brainard S, & Gillmore, G. (1999). National WEPAN pilot climate survey exploring the environment for undergraduate engineering students. *1999 International Symposium on Technology and Society - Women and Technology: Historical, Societal, and Professional Perspectives,* New Brunswick, NJ, United States. https://ieeexplore.ieee.org/document/787310

Molina-Gaudo, P., Baldassarri, S., Villarroya-Gaudo, M., & Cerezo, E. (2010). Perception and Intention in Relation to Engineering: A Gendered Study Based on a One-Day Outreach Activity. *IEEE Transactions on Education 53* (1), 61-70.

Moskal, B. & Skokan, C. (2011). Supporting the K-12 classroom through university outreach. *Journal of Higher Education Outreach and Engagement, 15*(1) pp. 53-75.

Nadelson, L.S. & Callahan, J. (2011). A comparison of two engineering outreach programs for adolescents. *Journal of STEM Education, 12*(1-2), pp. 43-54.

National Academy of Engineering Committee on Public Understanding of Engineering Messages (2008), *Changing the conversation: Messages for improving public understanding of*

engineering, Washington, DC: The National Academies Press. https://doi.org/10.17226/12187

National Center for Education Statistics (2018). Number and percentage distribution of science, technology, engineering, and mathematics (STEM) degrees/certificates conferred by postsecondary institutions, by race/ethnicity, level of degree/certificate, and sex of student: 2008-09 through 2017-18. Retrieved from https://nces.ed.gov/programs/digest/d19/tables/dt19_318.45.asp

Orsak, G.C. (2003). Guest editorial. K-12: Engineering's new frontier. *IEEE Transactions on Education 46* (2), 209-210.

Padilla, J.L., & Leighton, J. P. (2017). *Cognitive interviewing and think aloud methods.* In B. D. Zumbo & A. M. Hubley (Eds.), *Social indicators research series: Vol. 69. Understanding and investigating response processes in validation research* (p. 211–228). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-56129-5_12

Salkind, N.J. (2010). *Encyclopedia of research design*. SAGE Publications.

Smaill, C.R. (2010). The implementation and evaluation of a university-based outreach laboratory program in electrical engineering. *IEEE Transactions on Education*, *53*(1), pp. 12-17.

University of Connecticut (2022). University of Connecticut Engineering Ambassadors. Website: https://ambassadors.engr.uconn.edu

University of Nebraska (2018). Engineering Ambassadors Exit Interviews. Personal Communication.

Watkins, M.W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), pp. 219-246.

VITA

Melissa Gayle Kuhn

Old Dominion University
Department of Educational Foundations and Leadership
1226 W. 43rd St., Norfolk, VA 23508

EDUCATION

Ph.D. in Education, Concentration in Educational Psychology & Program Evaluation
Old Dominion University, Norfolk VA December 2022

M.A. in Teaching, PK-6
Virginia Commonwealth University, Richmond, VA June 2015

B.A. in Early and Elementary Education
Virginia Commonwealth University, Richmond, VA May 2015

PUBLICATIONS

Validation of the Ambassador Questionnaire for Undergraduate Students Conducting Engineering Outreach, 2022.
M.G. Kuhn, J.K. Garner, & S. Chappell Moots, International Journal of Engineering Education.

Invention Education as a Context for Children's Identity Exploration, 2021.
J.K. Garner, E. Matheny, A. Rutledge, & **M.G. Kuhn**, Journal of STEM Outreach.

PROFESSIONAL EXPERIENCE

Education Specialist, The Center for Educational Partnerships, Norfolk VA. June 2020 – Present.
Graduate Teaching Assistant Instructor, Educational Foundations and Leadership Department, Old Dominion University, Norfolk VA. August 2019 – August 2020.
Graduate Assistant Educational Program Manager, Batten College of Engineering and Technology, Old Dominion University, Norfolk VA. August 2017- August 2019.
4th Grade Teacher, Richmond Public Schools, Richmond, VA. June 2015 – June 2017.