University of Kentucky

UKnowledge

Theses and Dissertations--Science, Technology, Engineering, and Mathematics (STEM) Education

Science, Technology, Engineering, and Mathematics (STEM) Education

2014

MEASURING POST-SECONDARY STEM MAJORS' ENGAGEMENT IN SUSTAINABILITY: THE CREATION, ASSESSMENT, AND VALIDATION OF AN INSTRUMENT FOR SUSTAINABILITY CURRICULA EVALUATION

David L. Little II
University of Kentucky, d.l.little80@gmail.com

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Little, David L. II, "MEASURING POST-SECONDARY STEM MAJORS' ENGAGEMENT IN SUSTAINABILITY: THE CREATION, ASSESSMENT, AND VALIDATION OF AN INSTRUMENT FOR SUSTAINABILITY CURRICULA EVALUATION" (2014). *Theses and Dissertations--Science, Technology, Engineering, and Mathematics (STEM) Education*. 3.
https://uknowledge.uky.edu/stem_etds/3

This Doctoral Dissertation is brought to you for free and open access by the Science, Technology, Engineering, and Mathematics (STEM) Education at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Science, Technology, Engineering, and Mathematics (STEM) Education by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

David L. Little II, Student

Dr. Margaret Mohr-Schroeder, Major Professor

Dr. Robert Shapiro, Director of Graduate Studies

MEASURING POST-SECONDARY STEM MAJORS' ENGAGEMENT IN SUSTAINABILITY: THE CREATION, ASSESSMENT, AND VALIDATION OF AN INSTRUMENT FOR SUSTAINABILITY CURRICULA EVALUATION

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Education
at the University of Kentucky

By
David L. Little II

Lexington, Kentucky

Director: Dr. Margaret Mohr-Schroeder, Associate Professor of Mathematics Education

Lexington, Kentucky

2014

Copyright © David L. Little II 2014

ABSTRACT OF DISSERTATION

MEASURING POST-SECONDARY STEM MAJORS' ENGAGEMENT IN SUSTAINABILITY: THE CREATION, ASSESSMENT, AND VALIDATION OF AN INSTRUMENT FOR SUSTAINABILITY CURRICULA EVALUATION

Ongoing changes in values, pedagogy, and curriculum concerning sustainability education necessitate that strong curricular elements are identified in sustainability education. However, quantitative research in sustainability education is largely undeveloped or relies on outdated instruments. In part, this is because no widespread quantitative instrument for measuring related educational outcomes has been developed for the field, though their development is pivotal for future efforts in sustainability education related to STEM majors.

This research study details the creation, evaluation, and validation of an instrument – the STEM Sustainability Engagement Instrument (STEMSEI) – designed to measure sustainability engagement in post-secondary STEM majors. The study was conducted in three phases, using qualitative methods in phase 1, a concurrent mixed methods design in phase 2, and a sequential mixed methods design in phase 3. The STEMSEI was able to successfully predict statistically significant differences in the sample (n= 1017) that were predicted by prior research in environmental education. The STEMSEI also revealed statistically significant differences between STEM majors' sustainability engagement with a large effect size ($.203 \leq \eta^2 \leq .211$). As hypothesized, statistically significant differences were found on the environmental scales across gender and present religion. With respect to gender, self-perceived measures of emotional engagement with environmental sustainability was higher with females while males had higher measures in cognitive engagement with respect to knowing information related to environmental sustainability. With respect to present religion, self-perceived measures of general engagement and emotional engagement in environmental sustainability were higher for non-Christians as compared to Christians. On the economic scales, statistically significant differences were found across gender. Specifically, measures of males' self-perceived cognitive engagement in knowing information related to economic sustainability were greater than those of females. Future research should establish the generalizability of these results and further test the validity of the STEMSEI.

KEYWORDS: Sustainability, STEM, Education, Post-Secondary, Engagement

David L. Little II

August 7, 2014

MEASURING POST-SECONDARY STEM MAJORS' ENGAGEMENT IN SUSTAINABILITY: THE CREATION, ASSESSMENT, AND VALIDATION OF AN INSTRUMENT FOR SUSTAINABILITY CURRICULA EVALUATION

By

David L. Little II

Dr. Margaret J. Mohr-Schroeder
Director of Dissertation

Dr. Robert Shapiro
Director of Graduate Studies

August 7, 2014

To those who have walked with me on this journey, bless you and thank you.

To those who have walked with me in spirit, keep smiling down.

Onwards and upwards now to brighter days filled with love and awesomeness!

ACKNOWLEDGMENTS

To the many, many professors that helped me achieve this goal, I give my deepest heartfelt thank you. Specifically to Dr. Dusan Sekulic, Dr. Fazleenza Badurdeen, Dr. Gregory Luhan, and Dr. Leslie Vincent, thank you for your help teaching me about sustainability and helping guide me to this point; your time, effort, and feedback was pivotal to the success of this study and to the future fruits of this labor. For challenging me to push harder and make the most of the challenging project I chose, my sincere thanks to Dr. Jennifer Wilhelm and Dr. Bruce Walcott; I was able to accomplish things I thought impossible thanks to your thought provoking questions. To Dr. Michael Toland, your patience and time have been plentiful; I would not have been able to succeed in these last two years without your assistance. Finally, to Dr. Margaret Mohr-Schroeder, you have invested more than just time and energy in me; without your endeavors on my behalf, I would not have been able to make it to this point in so many ways. Thank you for everything, truly; I would not be the person or scholar I am today without your influence.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iii

LIST OF TABLES vii

LIST OF FIGURES ix

Chapter I Introduction 1
Statement of the Problem 1
Purpose of the Study and Research Questions 2
Significance of the Study 3
Theoretical Framework 3
Definition of Terms 6
Expected Outcomes for the STEMSEI 7
Expected outcomes across sustainability engagement. 8
Expected outcomes along economic engagement. 10
Expected outcomes along environmental engagement. 10
Expected outcomes along social engagement 11
Assumptions 11
Limitations of the Study 11
Purpose of the Instrument 11
Limitations of the Instrument 12
Considerations of Researcher Bias 12
Organization of the Study 13

Chapter II Review of Literature 15
Sustainability Engagement and the Need for Sustainability Education Measures 15
Introduction 15
Current quantitative assessment of sustainability education efforts. 16
Moving forward with quantitative assessment of sustainability education efforts... 18
A sustainability engagement framework conducive to measurement 20
Sustainability engagement: A sustainability competency useful and worth measuring. 23
A quantifiable definition of sustainability 24
Introduction. 24
A base definition for sustainability. 25
Quantifying environmental sustainability. 27
Quantifying economic sustainability. 30
Quantifying social sustainability 31
The quantifiable definition of sustainability. 32
Limitations of the quantifiable definition of sustainability 33
Theoretical framework overview. 34
Expected Outcomes Across the Literature in Sustainability Related Measures 35

Chapter III Methodology 37
Phase 1: Instrument Planning..... 40
Phase 1, step 1: Development of purpose, target population, and domains..... 40
Phase 1, step 2: Literature review and expert panel review..... 41
Phase 1, step 3: Objectives and item format. 46
Phase 2: Instrument Construction 48
Phase 2, step 4: Instrument blueprint and item writing. 48
Phase 2, step 5 and step 6: Content validation and item piloting. 51
Phase 3: Instrument Evaluation 59
Phase 3, step 7a: Preparing for instrument piloting. 59
Phase 3, step 7b: Pilot administration. 61
Phase 3, step 8a: Calculate reliability. 65
Phase 3, step 8b: Item analysis. 65
Synopsis 85

Chapter IV Results 86
Phase 1: Instrument Planning Results 86
Phase 1, step 1 results: STEMSEI purpose, target population, and domains. 86
Phase 1, step 2 results: Literature review and expert panel review. 88
Phase 1, step 3 results: Objectives and item format..... 90
Phase 2: Instrument Construction 90
Phase 2, step 4 results: Instrument blueprint and item writing. 90
Phase 2, step 5 and step 6 results: Content validation and item piloting. 92
Phase 3: Instrument Evaluation 96
Phase 3, step 7, round 1 results: Preparing for instrument piloting and pilot administration. 97
Phase 3, step 8a, round 1 results: Calculate reliability. 97
Phase 3, step 8b, round 1 results: Item analysis..... 97
Phase 3, step 9, round 1 results: Qualitative piloting in the sample. 114
Phase 3, step 7, round 2 results: Preparing for instrument piloting and pilot administration. 116
Phase 3, step 8a, round 2 results: Calculate reliability. 117
Phase 3, step 8b, round 2 results: Item analysis..... 117

Chapter V Discussion, Implications, and Conclusions 146
Hypotheses Testing 146
Research question 1. 146
Research question 2. 147
Research question 3. 147
Unexpected result..... 148
Hypothesis testing conclusions with respect to validity. 148
Considerations in STEMSEI Development, Evaluation, and Validation 149
Refining the STEMSEI 151
Implications for Sustainability Engagement in Post-secondary STEM Students 153
Differences in post-secondary STEM student sustainability engagement..... 153
Sustainability engagement invariance across post-secondary STEM students..... 156

Levels of sustainability engagement across STEM majors. 157
Future Uses of the STEMSEI 159
Final Implications 160
Recommendations..... 160

Appendices
APPENDIX A National Science Foundation STEM Majors 162
APPENDIX B Phase 1: Content Specialist Semi-Structured Interview Protocol..... 166
APPENDIX C Phase 2: Content Specialist Semi-Structured Interview Protocol..... 167
APPENDIX D Phase 2: Student Cognitive Interview Protocol 168
APPENDIX E Faculty Liaison Email 169
APPENDIX F Study Invitation Email..... 170
APPENDIX G Phase 3: Student Interview Protocol..... 172
APPENDIX H Phase 3, Round 1: Initial STEMSEI Instrument..... 173
APPENDIX I Phase 3, Round 2: Final STEMSEI Instrument..... 179

REFERENCES 186

VITA.....199

LIST OF TABLES

Table 2.1: Definition of Sustainability through Rates of Consumption and Renewal…..28
Table 2.2: Delineations of Sustainability Engagement as Proposed by the STEMSEI….35
Table 3.1: Content Analysis Codes for Content Expert Interviews…….…….…….46
Table 3.2: Developed Objectives for the STEMSEI……..47
Table 3.3: STEMSEI Blueprint……..……49
Table 3.4: Items to be Developed Based on Recommendations in the Literature…….53
Table 3.5: Demographic Information for First and Second Piloting Round Participants..63
Table 4.1: Developed Objectives for the STEMSEI…… .91
Table 4.2: STEMSEI Blueprint……..92
Table 4.3: EFA's for Emotional Engagement and Cognitive Engagement Items……..…98
Table 4.4: EFA's for Economic, Environmental, and Social Items……..…….99
Table 4.5: EFA Factor Structures for Economic and Environmental Items……..…..100
Table 4.6: Corresponding Bifactor Model Extractions for Economic and Environmental Items……..……..……101
Table 4.7: Confirmatory Factor Analysis and Confirmatory Bifactor Analysis Results for Validation Sample……..……..……101
Table 4.8: GPCM, GRM, and bifactor IRT models AIC statistics and AIC Difference Statistics…….103
Table 4.9: Initial Bifactor IRT Model Parameters for Economic Scale. ..…….…….104
Table 4.10, Initial Bifactor IRT Model Parameters for Environmental Scale…...……..105
Table 4.11, Phase 3, Round 1: Differential Item Functioning Tests for Economic Items……106-107
Table 4.12, Phase 3, Round 1: Initial Differential Item Functioning Tests for Environmental Items…….108-109
Table 4.13, Phase 3, Round 1: Follow-up Differential Item Functioning Tests for Environmental Items……...…….109
Table 4.14, Standard Error of Estimate (SEE) and Marginal Reliability Estimates for STEMSEI Scales…….….110
Table 4.15, Correlations of Person Scores (θ) Using Two Item Parameter Sets Across the Validation Sample……..……..111
Table 4.16, One-way ANOVA’s for First Round Economic and Environmental General Scale Scores……...…….114
Table 4.17, Final Bifactor Extractions for the Economic and Environmental Scales……..…….118
Table 4.18, Factor Patterns of an Exploratory Factor Analysis of the Social Items…....119
Table 4.19, Corresponding Bifactor Model Extraction Along the Social Items……..119
Table 4.20, GPCM, GRM, and Bifactor IRT Models AIC Statistics and AIC Difference Statistics…….……..120
Table 4.21, Final Bifactor IRT Model Parameters for Economic Scale……...…122-123
Table 4.22, Final Bifactor IRT Model Parameters for Environmental Scale……...123-124
Table 4.23, Final Bifactor IRT Model Parameters for Social Scale…….125-126
Table 4.24, Phase 3, Round 2: Differential Item Functioning Tests for Economic Items……..…..127-130

Table 4.25, Phase 3, Round 2: Differential Item Functioning Tests for Environmental Items……...……..131-134
Table 4.26, Phase 3, Round 2: Differential Item Functioning Tests for Social Items…….…….135-138
Table 4.27, Round 2 Standard Error of Estimate (SEE) and Marginal Reliability Estimates for STEMSEI Scales…….……..139
Table 4.28, Non-factorial Design MANOVA Analyses for STEMSEI Scale Scores …….....…….140
Table 4.29, Post-hoc ANOVA Analyses for STEMSEI Scale Scores with Benjamini-Hochberg correction.....141-142
Table 4.30, Post-hoc t-test Analyses for STEMSEI Scale Scores Identifying Sub-group with Higher Average Scores and Cohen's D …….....…….143
Table 4.31, Pearson's Correlations Among STEMSEI Scores ...……...….145

LIST OF FIGURES

Figure 2.1. The sustainability engagement framework .....……...……22
Figure 3.1. Study design of phase 1, instrument planning……...…….…40
Figure 3.2. Qualitative study design for steps 2b and 2c of phase 1…...……..45
Figure 3.3. Study design of phase 2, instrument construction……..……..49
Figure 3.4. Concurrent mixed methods design for content validation
and item piloting……52
Figure 3.5. Study design of phase 3, instrument evaluation…….…….59
Figure 3.6. Sequential mixed-methods design for STEMSEI piloting…….…….66
Figure 4.1. Study design of phase 1, instrument planning…….……87
Figure 4.2. Study design of phase 2, instrument construction……..91
Figure 4.3. Study design of phase 3, instrument evaluation……...……..96
Figure 4.4. Empirical Plots and Option Response Functions (ORF's)……112

Chapter I

Introduction

The concept of sustainability has grown and evolved over the past two decades, yielding substantial changes in how we view the connections between the economy, the environment, and our social structures. Through that evolution, the breadth and concerns of sustainability have been widely debated while no clear definition for sustainability has surfaced (Djordjevic & Cotton, 2011; Sekulic, 2011). At the same time, the importance and momentum of sustainability grows as our understanding of this concept evolves from new knowledge. While many post-secondary institutions have fostered both institutional change and academic reform to support sustainability, there is a lack of quantitative methods to assess these changes. Central to this study, the academic goals of such curricular reforms often lack proper instruments to assess student growth and change due to curricular intervention.

Statement of the Problem

As sustainability education has gained traction in post-secondary environments, the creation of sustainability courses and curricula at the post-secondary level has increased across the globe. However, as we draw near the end of the Decade of Education for Sustainable Development (United Nations Educational, Scientific and Cultural Organization, 2005), key questions still loom about how post-secondary sustainability curricula should be approached. Across the literature, there is a general theme of the disparity between knowledge for sustainability and action towards sustainability (see Christensen, Thrane, Jorgensen, & Lehmann, 2009; Davis, Edmister, Sullivan, & West, 2003; Hodson, 2003; O'Connell, Potter, Curthoys, Dyment, & Cuthbertson, 2005; Thapa, 1999; Van Kerkhoff & Lebel, 2006; Wright, Ironside, & Gwynn-Jones, 2009). This goal is especially important at the post-secondary level since recent and soon-to-be graduates from these institutions will lead and shape developments in sustainability for generations to come. However, if the disparity between knowledge for sustainability and action towards sustainability is to be bridged, we must identify the source of this gap and remedy it with curricular revisions and education research. In the current literature, instrumentation currently used for this purpose is outdated and does not

adequately measure the constructs of interest (e.g., Mann, Harraway, Broughton-Ansin, Deaker, & Shephard, 2013; Schneiderman & Freihoefer, 2012).

Purpose of the Study and Research Questions

The purpose of this three-phase, multi-methods study was to create, evaluate, and validate an instrument – the STEM Sustainability Engagement Instrument, or STEMSEI – measuring sustainability engagement in post-secondary science, technology, engineering, and mathematics (STEM) majors. The results of this study were meant to (1) provide post-secondary sustainability educators an instrument appropriate to assess sustainability engagement in post-secondary STEM students, (2) to provide means to establish norms for sustainability engagement for reference and research, and (3) to inform the development of future sustainability education instruments that may measure similar or related constructs of interest.

From the sustainability educator perspective, the STEMSEI would allow faculty to assess to what degree sustainability course interventions affect post-secondary STEM students' sustainability engagement by comparing, through statistical analyses, pre- and post-course measures of student sustainability engagement (Badurdeen et al., 2013). Moreover, the STEMSEI allows faculty members to assess sustainability engagement across various sustainability issues.

This study utilized an iterative instrument development process to develop the STEMSEI. In this process, an overarching research question was developed for each of the three main phases of this process. Specifically, the following questions were examined:

1. To what extent can a convergent theoretical framework for sustainability engagement in post-secondary STEM students be achieved between sustainability experts across the STEM disciplines?
2. To what extent can items that measure unique types of sustainability engagement and can be interpreted by post-secondary sustainability educators and post-secondary STEM students be developed across the STEM disciplines?
3. To what extent does the STEMSEI produce interpretable and useable/meaningful results with respect to sustainability education at the post-secondary level?

Significance of the Study

Several quantitative indicators for various aspects of sustainability have been developed (see United Nations Division for Sustainable Development, 2001). Quantitative indicators for sustainability help identify progress towards sustainability and sound early warnings for sustainability issues related to economic, environmental, or social factors (Hrebicek et al., 2013). However, to the researcher's knowledge, no quantitative indicators for measuring progress towards outcomes in sustainability education have been developed. The STEMSEI produces a quantitative measure of one such sustainability education outcome: sustainability engagement. With respect to sustainability curricula evaluation, the STEMSEI will allow faculty to assess the degree to which sustainability course interventions affect sustainability engagement in post-secondary STEM students. Such instruments could then be used to identify research-based instructional strategies that promote sustainability engagement in post-secondary STEM students, which the researcher posits would be of significant use in the field. Moreover, since some have argued that generalizable sustainability instruments are not possible (see Shriberg, 2002), this study offers proof that such sustainability instruments are possible and the methodology of this study may help inform future instrument development for other sustainability education outcomes.

Theoretical Framework

A theoretical framework for sustainability engagement was necessary to construct an instrument purporting to measure such a construct (Benson & Clark, 1982). The sustainability engagement framework posits that sustainability engagement can be comprehensively and distinctly described through three types of engagement: emotional, cognitive, and behavioral sustainability engagement.

The sustainability engagement framework was developed from a psychological framework for school engagement from Fredricks, Blumenfeld, and Paris' (2004). Fredricks et al. (2004) contend that school engagement consists of three dimensions: (1) behavioral engagement, (2) emotional engagement, and (3) cognitive engagement, and summarize as below:

> Behavioral engagement draws on the idea of participation; it includes involvement in academic and social or extracurricular activities and is considered crucial for achieving positive academic outcomes and preventing dropping out.

> Emotional engagement encompasses positive and negative reactions to teachers, classmates, academics, and school and is presumed to create ties to an institution and influence willingness to do the work. Finally, cognitive engagement draws on the idea of investment; it incorporates thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills. (p. 60)

A similar framework for sustainability engagement was developed. The Sustainability Engagement Framework is as follows:

- *Emotional sustainability engagement* is defined having/expressing any preference/opinion that aligns with the goals of sustainability. Examples of *emotional sustainability engagement* include (but are not limited to):
    - endorsement of opinions and/or preferences that prioritize sustainability;
    - comfort in challenging social normative assumptions counter to sustainability.
- *Cognitive Sustainability Engagement* is defined participating in any fact- or knowledge-driven mental process that aligns with the goals of sustainability. Examples of *cognitive sustainability engagement* include (but are not limited to):
    - investing the mental effort to comprehending sustainability;
    - knowing/comprehending complex ideas of sustainability;
    - knowing/refining the skills necessary to be sustainable;
    - problem solving or developing solutions for sustainability.
- *Behavioral Sustainability Engagement* is defined as participating in any action that aligns with the goals of sustainability, including fostering emotional, cognitive, or behavioral sustainability engagement in others. Examples of *behavioral sustainability engagement* include (but are not limited to):
    - implementing plans to live more sustainability at home (reduce consumption, use renewable resources/products in place of non-renewable resources/products, etc.);
    - implementing plans at work to increase the sustainability of professional environments;
    - showing/explaining through data-driven processes why sustainability is a necessity (to personal and/or professional contacts).

Emotional sustainability engagement, cognitive sustainability engagement, and behavioral sustainability engagement were theoretically viewed as unique and distinguishable from one another. The researcher argues that the sustainability engagement framework used in this study is (1) integrative and comprehensive over sustainability pedagogy, (2) accounts for the structure of the sustainability education classroom, and (3) extends to describe the function and practice of sustainability engagement by STEM practitioners in the field. Details are provided in Chapter II.

Since the Triple Bottom Line framework is sometimes criticized for valuing one sustainability domain over another (see Marshall & Toffel, 2005), minimal quantifiable criteria for sustainable states were defined for each domain of sustainability (economic, environmental, and social). This was done so that in item development for the STEMSEI, items could be written using language reflective of these minimal quantifiable sustainability states. Criteria for each of these minimal quantifiable sustainability states are given in the next section.

A quantifiable definition of sustainability was also developed to ensure that all aspects of sustainability engagement were considered from a content perspective (i.e., with respect to sustainability itself). Since there is no agreed upon definition for sustainability (Djordjevic & Cotton, 2011; Sekulic, 2011; White, 2013), such a framework for sustainability was developed from across the literature. The Triple Bottom Line framework (see Marshall & Toffel, 2005) posits that sustainability is composed of three domains: economic, environmental, and social. In effect, this means any issue related to sustainability may be regarded as an issue related to economic, environmental, and/or social concerns (Meadows, 2008). Many sustainability issues have concerns related in all three domains simultaneously (see Meadows, 2008; World Commission on Environment and Development, 1987).

The quantifiable definition of sustainability defines a development (commercial, industrial, or residential) to be sustainable if it meets the minimum level of performances in the following three areas of concern:

(1) Environmental Domain:
   a. renewable resources such as fish, soil, and groundwater must be used no faster than the rate at which they regenerate (Smith, 2004a, p. 1)
   b. nonrenewable resources such as minerals and fossil fuels must be used no faster than renewable substitutes for them can be put into place (Smith, 2004a, p. 1)
   c. pollution and wastes must be emitted no faster than natural systems can absorb them, recycle them, or render them harmless (Smith, 2004a, p. 1)

(2) Economic Domain: that the generalized production capacity of an economy is maintained intact, such as to enable at least constant consumption per capita through time (inspired by Solow, 1974, 1986, as cited in Hediger, 2006, p. 362)

(3) Social Domain: at least basic human needs (clean water for consumption and use, food, shelter, clothing, healthcare, and education) are met worldwide (inspired by Nattrass & Altomare, 1999, p. 23)

In effect, the quantifiable definition of sustainability delineates sustainability issues as related to economic, environmental, and/or social issues. That is, issues related to sustainability may be uniquely aligned to one of these domains, may be a combination of any two of these domains, or may combine all three domains.

Definition of Terms

The following terms are defined for this study. An in-depth review of the origins and motivations of these definitions is provided in Chapter II.

Emotional sustainability engagement is defined having/expressing any preference/opinion that aligns with the goals of sustainability. Examples of emotional sustainability engagement include (but are not limited to):

- endorsement of opinions and/or preferences that prioritize sustainability;
- comfort in challenging social normative assumptions counter to sustainability.

Cognitive sustainability engagement is defined participating in any fact- or knowledge-driven mental process that aligns with the goals of sustainability. Examples of cognitive sustainability engagement include (but are not limited to):

- investing the mental effort to comprehending sustainability;
- knowing/comprehending complex ideas of sustainability;
- knowing/refining the skills necessary to be sustainable;
- problem solving or developing solutions for sustainability.

Behavioral sustainability engagement is defined as participating in any action that aligns with the goals of sustainability, including fostering emotional, cognitive, or behavioral sustainability engagement in others. Examples of behavioral sustainability engagement include (but are not limited to):

- implementing plans to live more sustainability at home (reduce consumption, use renewable resources/products in place of non-renewable resources/products, etc.);
- implementing plans at work to increase the sustainability of professional environments;
- showing/explaining through data-driven processes why sustainability is a necessity (to personal and/or professional contacts).

Sustainable development or *sustainability* – a development (commercial, industrial, or residential) to be sustainable if it meets the minimum level of performances in the following three areas of concern:

(1) Environmental domain:
   a. renewable resources such as fish, soil, and groundwater must be used no faster than the rate at which they regenerate (Smith, 2004a, p. 1)
   b. nonrenewable resources such as minerals and fossil fuels must be used no faster than renewable substitutes for them can be put into place (Smith, 2004a, p. 1)
   c. pollution and wastes must be emitted no faster than natural systems can absorb them, recycle them, or render them harmless (Smith, 2004a, p. 1)

(2) Economic domain: that the generalized production capacity of an economy is maintained intact, such as to enable at least constant consumption per capita through time (Solow, 1974, 1986)

(3) Social domain: at least basic human needs (clean water for consumption and use, food, shelter, clothing, healthcare, and education) are met worldwide (Nattrass & Altomare, 1999, p. 23)

Development sample – a portion of the sample gathered in phase 3 that was used in initial quantitative analyses to assess fit of STEMSEI results to those expected from the theoretical framework developed in phase 1 and phase 2 of this study.

Validation sample – a portion of the sample gathered in phase 3 that was used in quantitative analyses following those performed with the development sample (see above); these analyses assessed fit of the STEMSEI results to those expected from the theoretical framework developed in phase 1 and phase 2 of this study.

STEM major - any major that aligns with the National Science Foundation's definition of STEM disciplines (see Appendix A).

Expected Outcomes for the STEMSEI

An instrument like the STEMSEI has not been created and utilized in the field of sustainability education to the knowledge of the researcher. Due to this, outcomes for such an instrument should be considered carefully. For example, would certain members of the population have higher scores on economic-related measures when compared to all other members of the population? Would the factor structure of responses match that of the theoretical framework used to develop the STEMSEI? Developing hypotheses such as this would help provide validity tests for the STEMSEI. While there were no such

results in the sustainability education literature to the researcher's knowledge, results from prior research in related fields could inform hypotheses concerning expected results from the STEMSEI.

To do this, the researcher explored literature related to quantitative measures of affective constructs (i.e., constructs related to personality, behavior, preferences, etc., but not measurements of content knowledge) in the economic, environmental, and social domains of sustainability. Prior literature concerning environmental attitudes related to sustainability seemed common. However, the opposite was true for the economic and social domains of sustainability; there seemed to be a gap in the literature on prior outcomes for economic or social measures related to sustainability, which may be due to the lack of quantitative instruments in the field. When possible, prior literature related to the sustainability domains was put in context of an engagement type (emotional, cognitive, and/or behavioral) to further strengthen the validity assessments.

For reader ease, the variables that were identified for possible statistically significant differences are listed here for preemptive consideration: race, gender, religion (past and present beliefs), classification (undergraduate/graduate student), and STEM major.

Expected outcomes across sustainability engagement. One hypothesis was that the factor structure of responses to the STEMSEI would match those indicated by the theoretical framework in some fashion. Details of factor analysis techniques are provided in detail later. Two possible outcomes were considered. In terms of utility, a factor structure matching the sustainability domains (i.e., measures for economic, environmental, and social engagement) would be helpful for sustainability faculty and might be reflected in responses to the STEMSEI. However, post-secondary STEM students may not perceive sustainability in this fashion. It may be that there is a greater degree of model-fit when considering the STEMSEI across engagement type (i.e., measures for emotional, cognitive, and behavioral engagement). Both would be reasonable results with respect to the theoretical framework of the instrument.

If the factor structure of responses to the STEMSEI reflected the three domains of sustainability (i.e., measures for economic, environmental, and social engagement), statistically significant differences were expected on all scores (i.e., economic,

environmental, and social) when comparing across STEM majors (see Appendix A). These statistically significant differences were theorized due to differences in traditional disciplinary focuses. In terms of emotional engagement, it is expected that post-secondary STEM students will have more positive preferences/opinions towards sustainability content that aligns with their major. This is theorized to be due to the natural proclivity towards certain preferences/opinions that are naturally cultivated in certain fields. For example, geosciences majors are theoretically more engaged in environmental sustainability than other STEM majors because they are naturally taught and/or trained to advocate preferences/opinions that support environmental awareness and stewardship. In contrast, the social sciences majors are theorized to be more engaged in social sustainability than other STEM majors because of the social sciences focus on issues that affect society and naturally cultivate a climate of preferences/opinions towards social awareness and advocacy. Similar arguments can be made across the other STEM disciplines for various other sustainability outcomes.

On the other hand, if the factor structure of responses to the STEMSEI reflected the three types of engagement (i.e., emotional, cognitive, and behavioral), no statistically significant differences are expected across STEM majors. This is because in such a factor structure, emotional engagement, for example, would encompass economic, environmental, and social items. Due to the mixing of various sustainability content into each engagement type, person location estimates may become skewed in terms of disciplinary content.

In contrast to STEM majors, it was expected that no statistically significant differences would exist between undergraduate and graduate students. The researcher argues that this is because undergraduate and graduate programs tend not to be distinctive based on content knowledge or opinions/preferences expressed across these programs. Rather, they are distinctive in the depth and breadth of material covered. This additional depth and breadth in material should not statistically significantly change the measured levels of sustainability engagement in undergraduate and graduate students.

Hence, it is hypothesized that STEM majors will differ statistically significantly from one another on all measures produced by the STEMSEI while no statistically

significant differences are expected based on classification (undergraduate/graduate student)

Expected outcomes along economic engagement. No prior quantitative research with affective economic scales could be identified by the researcher that would have related to this work (i.e., work in economic sustainability). Therefore, no additional hypotheses were identified with respect to economic engagement.

Expected outcomes along environmental engagement. Prior research in environmental education outcomes revealed there might be multiple expected outcomes along the environmental scale scores.

It was hypothesized that there would be statistically significant differences in environmental engagement measures based on race, gender, and religious beliefs. In order to compensate for small subsamples due to race, this variable was recoded for all respondents to “students of color” or “white.” A growing body of literature has highlighted the salient differences between white students and students of color in post-secondary STEM environments (see Griffin, Perez II, Holmes, & Mayo, 2010; Palmer, Maramba, & Elon Dancy II, 2011), which is similar to the differences Johnson, Bowker, and Cordell (2004) found in environmental belief and behavior based on ethnic variation. Whites were also found to score higher on the New Environmental Paradigm (see Dunlap, Van Liere, Mertig, & Jones, 2000) when compared to African Americans and foreign-born Latinos (Johnson, Bowker, & Cordell, 2004). Similarly, since previous literature has posited significant differences for Christians along environmental measures (see Lalonde & Jackson, 2002; Thapa, 1999), both past religious exposure and present religious beliefs were recoded as either Christian or non-Christian. Questions concerning past and present religious practice were open-ended questions (i.e., responses were provided in a text box for each, separately), so responses were recoded into "Christian" (= 0) and "non-Christian" (= 1) categories. Those who responded as “Unitarian” were recoded as missing data since no clear delineation as “Christian” or “non-Christian” can be made for this belief perspective (see British Broadcasting Corporation, 2004). Similarly, those who responded as “Agnostic” were recoded as missing data and were not considered for analyses involving religion.

Expected outcomes along social engagement. As with the economic scales, no prior quantitative research with affective social scales could be identified by the researcher that would have related to this work (i.e., work in social sustainability). Therefore, no additional hypotheses were identified with respect to social engagement (i.e., no statistically significant differences were expected across race, gender, religion, or classification).

Assumptions

1. The expert panelists provided accurate information.
2. The post-secondary STEM participants provided both accurate information and responses.
3. The sustainability engagement constructs (emotional and cognitive) were a continuous latent trait that can be measured through ordinal responses.

Limitations of the Study

1. Development of items that measured behavioral sustainability engagement could not be included in this study due to limitations in time and resources.
2. While the study utilized data from post-secondary students majoring in STEM fields across the United States, calibration of item parameters by item response theory models were only estimates of population measures. This limits the extent to which generalizations with these measures can be made, especially over such a broad population as post-secondary STEM majors in the United States.
3. Some invitations to participate in the study were sent electronically via listservs at participating institutions. Response rate could only be estimated for these institutions.
4. Some institutions elected to invite only a fraction of their STEM majors.
5. At one participating institution where all applicable STEM majors were targeted, some departments elected not to send study invitations to their undergraduate and/or graduate students.
6. Multidimensional differential item functioning analyses were not able to be performed due to software limitations. In such cases, comparable unidimensional differential item functioning analyses were performed, though future studies should seek to employ analyses that are more appropriate.

Purpose of the Instrument

The purpose of the STEMSEI was to measure sustainability engagement across post-secondary STEM majors. Again, from the sustainability educator perspective, the STEMSEI would allow faculty to assess to what degree sustainability course interventions affect post-secondary STEM students' sustainability engagement by comparing, through statistical analyses, pre- and post-course measures of student

sustainability engagement (Badurdeen et al., 2013). Moreover, the STEMSEI allows faculty members to assess sustainability engagement across various sustainability issues.

Limitations of the Instrument

1. Behavioral sustainability engagement is a necessary component of the sustainability engagement framework, but was not developed/measured in this study due to limitations in resources and time. Hence, the STEMSEI does not measure behavioral sustainability engagement.
2. The STEMSEI was calibrated in this study with a sample of students almost exclusively from two universities located in the southern region of the United States. Due to this, the item parameters estimated in this study may not be accurate reflections of population item parameters for all post-secondary STEM students in the United States. The generalizability of item parameters for the STEMSEI should be assessed in future work.
3. Scores produced by the STEMSEI should not be used as measures of students' knowledge of sustainability for any grading purpose.
4. Scores produced by the STEMSEI should not be used as measures of an instructor's teaching effectiveness or for any other administrative evaluative decision/process.
5. The STEMSEI should not be used for the evaluation of sustainability curricula *unless* one of the curricular goals of such curriculum is to increase student sustainability engagement as previously defined. In such a case, this instrument should only be used to identify statistically significant changes in sustainability engagement and no other variable/construct.

Considerations of Researcher Bias

With respect to paradigm and worldview, aspects of researcher bias can influence research (McMillan & Schumacher, 2010). With respect to sustainability, the fact that there is no agreed upon definition for sustainability (Djordjevic & Cotton, 2011; Sekulic, 2011; White, 2013) may be seen as evidence of these multiple sustainability paradigms and worldviews. As such, the researcher's own paradigms and worldviews were considered to identify potential bias.

The researcher is the first college-educated member of his family and is originally from eastern Kentucky. Due to growing up in this region, the researcher was influenced from childhood by the “coal culture.” Though the researcher perceives coal as an important economic stimulant to the region, the researcher believes a more thorough analysis of the full life cycle of coal extracted from eastern Kentucky should be considered. For example, the researcher believes the long-term effects of coal extraction on ecological systems and human health need to be considered in light of the economic

stimulus coal brings to the area. While coal does “keep the lights on”, it comes at a price that we must consider on multiple levels. Due to these perspectives, the researcher advocates for alternative, renewable energy sources that may still offer an economic stimulus to this region of the country.

The researcher also participated as a student in a course funded by a National Science Funded Transforming Undergraduate Education in STEM that focused on sustainability and sustainability issues. The course was designed for business, design, education, and engineering undergraduate and graduate students (for further course description details, see Badurdeen et al., 2013). Due to participating in this course, the researcher became interested in sustainability education in STEM majors after noticing the differences in how other students in the course reacted to course material, instructional methods, and course content. Due to the lack of instrumentation to assess the curricular impacts in this course, the researcher became interested in developing a tool that could help with the evaluation of such sustainability curricula.

Organization of the Study

The goal of this study was to determine whether an instrument could be developed to measure sustainability engagement in post-secondary STEM students and if that instrument performed as expected within the population. The necessity of such an instrument was presented in Chapter I with the supporting theoretical framework developed from the literature following in Chapter II. A development and analysis plan, presented in Chapter III, was created to assess the instrument's validity and generalizability to the population. Results and findings of the analyses are presented in Chapter IV. Finally, Chapter V synthesizes the results and findings with prior sustainability education literature to (1) inform future use of the STEMSEI and (2) inform future instrument development for sustainability education assessment instruments. The findings of this study will be shared with sustainability educators to inform future instrument development and to disseminate the instrument for use in the field.

I. Introduction
II. Literature Review
III. Methodology
IV. Results

V. Discussion, Conclusions, and Implications

Copyright © David L. Little II 2014

Chapter II

Review of Literature

Chapter II contains a literature review on three major topics that helped shape this study. First, the need for sustainability education measures and a theoretical framework for sustainability engagement is presented. Second, definitions of sustainability from across multiple fields are synthesized to form a quantifiable definition for sustainability that was utilized for this study. Finally, outcomes from similar research were used to develop expected outcomes/hypotheses for the STEMSEI.

Sustainability Engagement and the Need for Sustainability Education Measures

This section presents the development of the sustainability engagement framework used in this study to delineate it into distinct and representative subdomains of engagement (emotional, cognitive, and behavioral) which then guided item development. This was important to ensure all aspects of sustainability engagement were accounted for in the theoretical framework. At the same time, the sustainability engagement framework had to be interpretable and meaningful for sustainability educators across STEM education if the STEMSEI is to be used in the field for evaluating sustainability curricula.

Introduction. As sustainability education has gained traction in post-secondary environments, it is unclear what tools are available for sustainability educators at the post-secondary level to assess the effectiveness of curricular interventions in closing the gap between knowledge and action for sustainability. Many sustainability education studies rely on qualitative methods (see Christensen, Thrane, Jorgensen, & Lehmann, 2009; Davis, Edmister, Sullivan, & West, 2003; O'Connell, Curthoys, Dyment, & Cuthbertson, 2005) while quantitative studies have relied on outdated instruments that do not measure constructs encompassing sustainability as a whole (see Mann, Harraway, Broughton-Ansin, Deaker, & Shephard, 2013; Schneiderman & Freihoefer, 2012). The latter are focused on here to understand the need for quantitative assessment instruments in sustainability education.

Critical assessment of evaluation tools for sustainability education is necessary when considering one of the goals of sustainability education at the post-secondary level: to engender knowledge and action for sustainability (United Nations Educational, Scientific and Cultural Organization, 2005). Given the general disparity between

knowledge for sustainability and action towards sustainability (Christensen, Thrane, Jorgensen, & Lehmann, 2009; Davis, Edmister, Sullivan, & West, 2003; Hodson, 2003; O'Connell, Curthoys, Dyment, & Cuthbertson, 2005; Thapa, 1999; Van Kerkhoff & Lebel, 2006; Wright, Ironside, & Swynn-Jones, 2009), curricular evaluation tools could be used to help identify instructional methods that help alleviate this gap.

Current quantitative assessment of sustainability education efforts. The fact that faculty across multiple institutions have assessed the curricular effectiveness of many post-secondary sustainability education efforts is not in question. On the contrary, several studies have looked at this issue (see Brundiers & Wiek, 2011; Christensen, Thrane, Jorgensen, & Lehmann, 2009; Mann, Harraway, Broughton-Ansin, Deaker, & Shephard, 2013; O'Connell, Curthoys, Dyment, & Cuthbertson, 2005; Schneiderman & Freihoefer, 2012). From the researcher's perspective, critical analysis of the research methods used in these studies reveals some flaws in study design, however.

Consider the New Ecological Paradigm (NEP) (Dunlap, Van Liere, Mertig, & Jones, 2000). The NEP has been used to research sustainability education, including identifying curricular needs of post-secondary students with respect to sustainability (see Mann, Harraway, Broughton-Ansin, Deaker, & Shephard, 2013) and determining effectiveness of instructional methods (see Schneiderman & Freihoefer, 2012). The NEP has a long history in the field of environmental education, designed to measure environmental concern in respondents (Dunlap, 2008). The 15 items of the NEP instrument all refer to environmental or ecological ideas and/or environmental/ecological-related situations (for full NEP instrument, see Schneiderman & Freihoefer, 2012). Moreover, the NEP is traditionally viewed as a multidimensional instrument (Cordano, Welcomer, & Scherer, 2003), with five environmental/ecological subdomains; (1) environmental limits to growth, (2) antianthropocentrism (from environmental/ecological contexts), (3) the fragility of nature's balance, (4) rejection of environmental exemptionalism, and (5) fear of ecological crises (Dunlap et al., 2000). Many of these dimensions have 4 or fewer items, which is less than the recommended 7 to 10 items for subdomains of an instrument (Bohrnstedt, 2010, p. 375). Moreover, despite the NEP's design and focus for measuring only environmental concern, it continues to be an instrument used to research sustainability education (see Mann et al.,

2013; Schneiderman & Freihoefer, 2012). This approach to research sustainability education solely through environmental indicators may be indicative of a misconception: that sustainability encompasses only environmental issues (Lemonick, 2009; McDonough & Braungart, 2002).

For example, Schneiderman and Freihoefer (2012) employed the NEP in a quasi-experimental design to "determine teaching method effectiveness" (p. 415) for enhancing interior design students' pro-environmental perceptions and environmental awareness in "ecological, economic, and sustainable" (p. 413) contexts. However, how can the NEP be used to assess teaching effectiveness in economic or sustainable contexts if those ideas are outside the contexts in which the instrument functions or was intended to measure (i.e., environmental perceptions and awareness)? The researcher argues that economic and sustainable contexts must be considered beyond environmental perceptions and awareness, though Schneiderman and Freihoefer's (2012) study focuses on environmental issues. While the study offered by Schneiderman and Freihoefer (2012) is one that is important and needed for sustainability education, the instrumentation employed to answer these questions was misaligned and insufficient for the task.

Mann, Harraway, Broughton-Ansin, Deaker, and Shephard (2013) used the NEP to produce thematic learning needs for sustainability across post-secondary students from seven departments using a statistical cluster-analysis. Mann et al. (2013) claim that knowing students' locations across thematic learning needs and sustainability perspectives could allow for a more tailored educational experience, theoretically producing more sustainability-oriented students. However, the NEP is not an instrument suited to measure anything except environmental concerns (Dunlap, 2008). Mann et al. (2013) do not define sustainability in their work, so it is unclear what domains beyond the environmental domain they consider part of sustainability. From the perspective of this research, such a view of sustainability is not encompassing of all the necessary components of sustainability, such as economic and social concerns. Since sustainability also includes social and economic domains (Bencze, Sperling, & Carter, 2011; Erdogan, 2010; Sekulic, 2011; Wainwright, 2010), how can the NEP be solely used to target learning needs for sustainability?

It is sometimes a misconception that sustainability is only about environmental concerns (Lemonick, 2009). The NEP instrument was not designed to measure sustainability as defined as a concept encompassing economic, environmental, and social concerns. The research questions that drove both of these studies were ones that should be pursued if we are to understand in what ways sustainability education can affect students' perceptions and actions towards sustainability. However, the instrumentation used to answer these research questions should match the context in which the questions were asked.

Moving forward with quantitative assessment of sustainability education efforts. For Schneiderman and Freihoefer's (2012) study, an instrument that measures economic and environmental perspectives related to sustainability would be more appropriate. This would have allowed the researchers to discern the impacts of the course on students' pro-environmental perspectives from economic contexts. Further, if an instrument were able to delineate measures along the environmental and economic domains of sustainability from various perspectives or contexts, this would have added new depth to Schneiderman and Freihoefer's (2012) study to contextualize and measure differences in their sample.

As for Mann et al.'s (2013) study, an instrument tailored towards measuring pro-sustainability perspectives from economic, environmental, or social contexts would have been more appropriate. This would allow departments and universities to create learning programs that are supportive of sustainability from the specific contexts (economic, environmental, and/or social) of their students' needs. Tailoring education efforts along such lines would also be easier to align students' needs with respect to content knowledge for sustainability to specific departments within post-secondary institutions. Additionally, different students will need different interventions based on their current knowledge (Vygotsky, 1978). Identifying differing student needs requires a more clear learning theory for sustainability education, which is absent in the literature as of this writing. Hence, even if an instrument were able to measure along the different domains of sustainability, it should also differentiate in varying student needs (i.e., students' zones of proximal development) (Vygotsky, 1978) to create entire programs surrounding

sustainability education. In the case of both of these research studies, such instruments do not currently exist.

Consider the instrument recommendations for Schneiderman and Freihoefer's (2012) and Mann et al.'s (2013) studies. If such instruments did exist, one important question that arises is how universal should such instruments be? The answer to this question would affect both instrument design and function (Linn, 2006). If such a universal instrument for researching sustainability education were developed, it would give faculty common frames of reference (i.e., common frames of measurement) to compare sustainability education research results. This could lead to the identification of best practices for sustainability education. However, in general, arguments against the development of universal tools for assessing sustainability in post-secondary institutions have been presented in the literature. For example, Shriberg (2002) argues that developing a universal sustainability assessment would be a "painstaking process, which would take longer than many stakeholders are willing to wait for results" (p. 165). The lack of contextual information in such universal sustainability assessments also poses a problem for development (Shriberg, 2002). While these concerns may be valid for creating universal assessments for sustainability in general, sustainability engagement may be beyond the purview of what Shriberg (2002) envisioned. Sustainability education may provide a much more structured and similar environment across all post-secondary institutions in terms of assessing curricular effectiveness, especially since assessing educational programs is common practice in post-secondary environments (see Astin & Antonio, 2012). These similarities across institutions and disciplines could be leveraged to develop a universal assessment framework for sustainability education across post-secondary institutions. This calls upon post-secondary educators to develop ways to get students engaged in sustainability and keep them engaged in sustainability outside of the classroom as well.

If such a framework could be rooted in identifying core competencies for sustainability knowledge and action, this would allow sustainability education researchers to alleviate the gap between knowledge and action for sustainability (see Christensen et al., 2009; Davis et al., 2003; Hodson, 2003; O'Connell et al., 2005; Thapa, 1999; Van Kerkhoff & Lebel, 2006; Wright et al., 2009). To continue to progress in sustainability,

post-secondary institutions must facilitate student involvement and buy-in (Brundiers & Wiek, 2011). A framework that delineates sustainability along the economic, environmental, and social domains would be helpful and conducive for answering the research questions driving both Mann et al.'s (2013) and Schneiderman and Freihoefer's (2012) studies. With respect to the STEM disciplines, delineations along such lines would also help pedagogically because of content knowledge alignment that could be matched to each of the sustainability domains (economic, environmental, and social). This implies that instruments based off of such frameworks should be able to differentiate between STEM majors. The next section provides such a framework that meets all of these criteria.

A sustainability engagement framework conducive to measurement. A lack of engagement in sustainability has been documented across the literature. Brown (2011) found that 6 out of 10 commercially employed engineers found sustainable measures and practices too complex to incorporate into their work. Businesses and post-secondary institutions sometimes find it difficult to transition from sustainability rhetoric to sustainability action (Christensen et al., 2009; Wright et al., 2009). Economic concerns are also often seen as more important than environmental concerns, which sometimes leads to a lack of engagement in environmental issues (Ameer & Othman, 2012; Anderson, 2009; Dyllick & Hockerts, 2002; Fairbrass & Zueva-Owens, 2011; Johnston, Everard, Santillo, & Robèrt, 2007; Norman & MacDonald, 2004). Finally, students in environmental courses often have a lack of transition from pro-environmental thoughts to pro-environmental action (Bencze, Sperling, & Carter, 2011; Hughes & Estes, 2005; Thapa, 1999). Though the lack of engagement in sustainability is documented, it still remains a question of how one would actually measure such a concept. Antecedently, this means that a framework for sustainability engagement must be developed that is conducive to measurement.

What exactly would sustainability engagement look like, though? A psychological framework for school engagement from Fredricks, Blumenfeld, and Paris (2004) provides a basis for a comparable sustainability engagement framework that is also quantifiable and measurable. Fredricks et al.'s (2004) school engagement framework

consists of three dimensions: (1) behavioral engagement, (2) emotional engagement, and (3) cognitive engagement, which are summarized below:

> "*Behavioral engagement* draws on the idea of participation; it includes involvement in academic and social or extracurricular activities and is considered crucial for achieving positive academic outcomes and preventing dropping out. *Emotional engagement* encompasses positive and negative reactions to teachers, classmates, academics, and school and is presumed to create ties to an institution and influence willingness to do the work. Finally, *cognitive engagement* draws on the idea of investment; it incorporates thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills" (p. 60).

For the needs of this study, sustainability engagement was considered in each of these forms. Correspondingly, below are the definitions for each type of sustainability engagement developed by the researcher.

Emotional sustainability engagement is defined as having/expressing any preference/opinion that aligns with the goals of sustainability. Examples of *emotional sustainability engagement* include (but are not limited to):

- endorsement of opinions and/or preferences that prioritize sustainability;
- comfort in challenging social normative assumptions counter to sustainability.

Cognitive Sustainability Engagement is defined as participating in any fact- or knowledge-driven mental process that aligns with the goals of sustainability. Examples of *cognitive sustainability engagement* include (but are not limited to):

- investing the mental effort to comprehending sustainability;
- knowing/comprehending complex ideas of sustainability;
- knowing/refining the skills necessary to be sustainable;
- problem solving or developing solutions for sustainability.

Behavioral Sustainability Engagement is defined as participating in any action that aligns with the goals of sustainability, including fostering emotional, cognitive, or behavioral sustainability engagement in others. Examples of *behavioral sustainability engagement* include (but are not limited to):

- implementing plans to live more sustainably at home (reduce consumption, use renewable resources/products in place of non-renewable resources/products, etc.);
- implementing plans at work to increase the sustainability of professional environments;
- showing/explaining through data-driven processes why sustainability is a necessity (to personal and/or professional contacts).

Based on these definitions, all three forms of sustainability engagement are necessary to describe individuals across the spectrum of sustainability engagement. For example, for individuals who see sustainability as a "waste of time" and a "manufactured concern", they would have no engagement in all three dimensions. On the other hand, an individual who sees the importance of sustainability but does not comprehend the complexities of it or the actions needed to achieve it would exhibit only emotional sustainability engagement. For individuals who actively seek to behave in sustainable ways, there would be engagement in all three dimensions. The three dimensions of sustainability engagement might interact in a sequential and nested fashion (see Figure 2.1). This implies that there may be dependencies of higher-tier engagements (behavioral) to lower-tier engagements (emotional). However, the work in this study is insufficient to test this hypothesis and should be considered in future research.

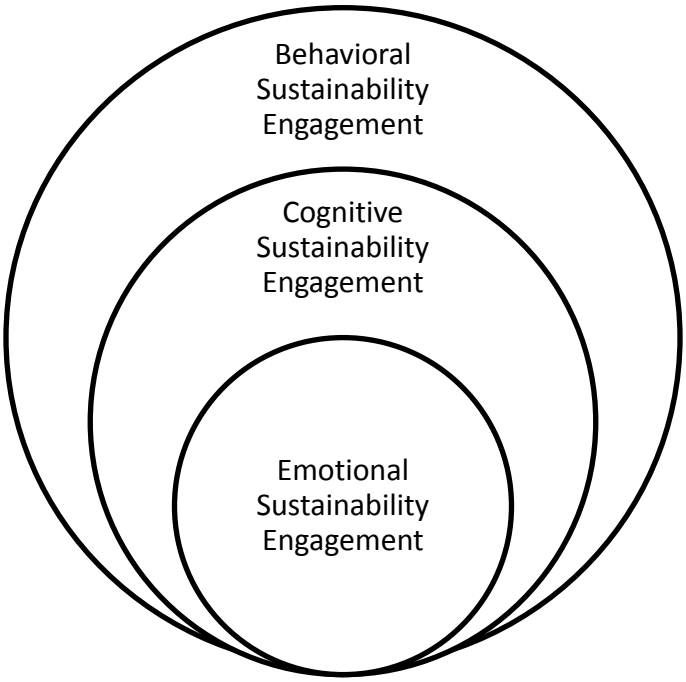

Figure 2.1. The sustainability engagement framework.

Sustainability engagement: A sustainability competency useful and worth measuring. Measuring post-secondary students' sustainability engagement along all three dimensions would be informative in many ways. Such measures could be used to assess for a connection between the different types of sustainability engagement, leading to a better understanding of if and how these engagement types build in individuals. Such measures could also be used to assess curricular efforts in increasing student sustainability engagement in particular contexts (i.e., along sustainability domains or along specific engagement types), or to identify curricular elements that lead to higher sustainability engagement. In a more general use, such measures could also be used to assess longitudinal national or global sustainability engagement. Also, recall that there is often a gap between pro-sustainability knowledge and pro-sustainability action Christensen, Thrane, Jorgensen, & Lehmann, 2009; Davis, Edmister, Sullivan, & West, 2003; Hodson, 2003; O'Connell, Curthoys, Dyment, & Cuthbertson, 2005; Thapa, 1999; Van Kerkhoff & Lebel, 2006; Wright, Ironside, & Swynn-Jones, 2009). Measures of knowledge for sustainability (i.e., cognitive engagement) and action for sustainability (i.e., behavioral engagement) could be used to identify such gaps in individuals and the possible causes for these gaps. The sustainability engagement framework can provide a setting for such measures. Even more important, if an instrument measuring along this framework were developed, it could then help to answer many of the important aforementioned questions in sustainability education.

As an example, by measuring sustainability engagement, sustainability education researchers could improve sustainability education efforts in many ways, such as tailoring sustainability curricula to student needs. By having an instrument that could assess student levels in each of these engagement types across the sustainability domains, one could replicate Mann et al.'s (2013) study to place students in the course most appropriate for their sustainability engagement deficiencies or sustainability content knowledge deficiencies. Course placements made in this way would greatly enhance the placements described by Mann et al. (2013) by addressing the needs of each student based on their current knowledge (Vygotsky, 1978). Moreover, Schneiderman and Freihoefer's (2012) study would have benefited from an instrument measuring across more than just environmental perspectives, as it is essential to consider economic and social dimensions

of sustainability as well (Bencze et al., 2011; Erdogan, 2010; Sekulic, 2011; Wainwright, 2010).

While there are certainly drawbacks to creating universal measures for various concepts related to sustainability education, there are also several advantages to doing so. As Shriberg (2002) points out, this will no doubt be a difficult and pain-staking process. However, universal assessment offers something unique that localized, individual assessment cannot: a collective insight into the national or global state of sustainability education. Sustainability educators need to focus on localized and universal assessments for sustainability education in contexts that are powerful, meaningful, and informing for current and future sustainability education efforts.

A quantifiable definition of sustainability

This section presents the development of a quantifiable definition of sustainability that was used in this study to guide item development. This was important because it established a standard for minimum sustainable state(s) across the various sustainability topics in which the population would be asked about. Simultaneously, this quantifiable definition of sustainability had to be interpretable and meaningful for sustainability educators across STEM education. As argued for in the "Sustainability Engagement and the Need for Sustainability Education Measures" section (see beginning of Chapter II), the delineation across the domains of economic, environmental, and social seemed to be most useful and practical for sustainability educators. Hence, that approach is developed from the literature.

Introduction. As we continue to understand the impact of human activity on global environments, both human and natural, our world perspectives change in order to accommodate the new understandings discovered. Understandings of the negative environmental effects faced in the 1970's paved the way for new political and societal views (Dunlap, Van Liere, Mertig, & Jones, 2000). From that time, Erdogan (2010) accounts for the evolution that Americans have experienced as we are now beginning to focus on environmental issues through the lens of sustainability. It seems that has caused some confusion on the topic, however. Sustainability is often thought to be mostly an environmental concept (Lemonick, 2009). While the environmental domain is essential when addressing sustainability, the topic encompasses much more than just the

environment (Lemonick, 2009; McDonough & Braungart, 2002), such as complex social and economic factors (Bencze, Sperling, & Carter, 2011; Erdogan, 2010; Sekulic, 2011; Wainwright, 2010).

Sustainability seeks to address many issues on a global scale. In truth, no one human being could understand sustainability in all its facets, as the issue and its subjects are much too vast. This and other factors contribute to sustainability being a transdisciplinary field (Le Grange, 2011). Moreover, the generalized concept of sustainability is a large and pervasive issue that currently defies definition (Djordjevic & Cotton, 2011; Sekulic, 2011). The breadth and pervasiveness makes the issue hard to focus and understand for most audiences. Moreover, a common definition of sustainability has yet to be identified (Djordjevic & Cotton, 2011; Sekulic, 2011; White, 2013), let alone a quantifiable one. This has led some authors to provide sustainability frameworks instead of definitions, as they have advocated that, given the nature of sustainability, this perspective of sustainability will be much more practical and effective (Gagnon, Leduc, & Savard, 2009; National Research Council, 1999; Quental, Lourenço, & da Silva, 2011; Talbot & Venkataraman, 2011).

With the many definitions and frameworks for sustainability that are offered across the field, careful consideration must be given to construct a definition for this study that (1) matches current perceptions of sustainability and is still interpretable and meaningful to sustainability educators while (2) still yet offering quantifiable sustainable states by which to measure post-secondary STEM students' sustainability engagement. Various definitions from the field are now presented that contributed to the quantifiable definition of sustainability that was used in this study.

A base definition for sustainability. Following the many attempts of the United Nations to address the environmental concerns of the 1960's and 1970's, a new initiative began in the early 1980's. The United Nations formed the World Commission on Environment and Development (WCED), chaired by Gro Harlem Brundtland. The WCED's (1987) report, "Our Common Future", now referred to as the Brundtland Report, defined sustainability through the concept of a sustainable development. The Brundtland Report asserted that a sustainable development is a "development that meets the needs of the present without compromising the ability of future generations to meet

their own needs" (Chapter 2, para. 1). The definition of sustainability the Brundtland Report gave became commonly referred to as the *Brundtland definition of sustainability*. Moreover, it became one of the most widely cited definitions of sustainability (Ameer & Othman, 2012; Bell & Morse, 2008; Borghesi & Vercelli, 2008; Duchin, 1996; Gagnon, Leduc, & Savard, 2009; Pezzoli, 1997; Quental, Lourenço, & da Silva, 2011; Ruttan, 1994).

The Brundtland Report also offered aspects of sustainability that were of great importance, such as equitable sharing and use of resources across the globe and global economic stimulus (see WCED, 1987). However, neither the Brundtland definition of sustainability nor the Brundtland report clearly identified what should be sustained or how sustainability should be measured (United Nations Economic Commission for Europe, 2005). However, it did delineate sustainability along economic, environmental, and social concepts (Basiago, 1995). These delineations were later reflected in metrics developed to quantify and assess sustainability (see United Nations Division for Sustainable Development, 2001) and those delineations are reflected in sustainability education research (e.g., Mann, Harraway, Broughton-Ansin, Deaker, & Shephard, 2013; Schneiderman & Freihoefer, 2012).

Though the Brundtland definition for sustainability may be among the most cited (Ameer & Othman, 2012; Bell & Morse, 2008; Borghesi & Vercelli, 2008; Duchin, 1996; Gagnon, Leduc, & Savard, 2009; Pezzoli, 1997; Quental, Lourenço, & da Silva, 2011; Ruttan, 1994), it can only serve to guide the direction of this research. It cannot provide a quantifiable sustainable state(s) by which to measure post-secondary STEM student's engagement with. Simply put, the Brundtland definition is an intuitive definition of sustainability that does not specify what is being sustained. While the Brundtland definition of sustainability does not specify sustainability in specific terms, it has been a driving force in the evolution of the field. From the perspective of this research, however, definitions and frameworks for sustainability are only informative if they offer a way to quantifiably understand sustainability. Frameworks and definitions for sustainability will now be presented that best provide a quantifiable approach to the three domains of sustainability: the environmental, economic, and social domains (see Basiago, 1995; Kiewiet & Vos, 2007).

The Triple Bottom Line framework (see Kiewiet & Vos, 2007) is a framework for sustainability that delineates sustainability along these lines. From the Triple Bottom Line perspective, "an organisation is considered sustainable if a certain minimum level of performance is attained [in each domain of sustainability]" (Kiewiet & Vos, 2007, p. 4). Moreover, it is expected that organizations utilizing this approach make decisions based on not only one domain, but all three; namely, profits should be considered in light of (1) environmental impacts that attaining those profits will cause and (2) social justice issues that may arise from such activities (Marshall & Toffel, 2005). Hence, the Triple Bottom Line framework views sustainability as three distinct domains (environmental, economic, and social) and seeks to balance them all separately (Kiewiet & Vos, 2007; Marshall & Toffel, 2005). In this sense, the Triple Bottom Line framework embodies all three domains of sustainability delineated in the Brundtland definition.

While this is often the predominant sustainability framework utilized in business/organizational sustainability (Kiewiet & Vos, 2007), there remains the unanswered question of how to balance the concerns of the economy, environment, and society within the Triple Bottom Line framework (Kiewiet & Vos, 2007; Norman & MacDonald, 2004; Smith & Sharicz, 2011). Namely, how does one quantify and measure what is to be sustained if those constructs are yet to be identified with respect to this framework? It offers a support of the Brundtland definition by considering the three domains of environmental, social and economic as core to achieving sustainability. However, a criticism of the Triple Bottom Line is that when some businesses apply the framework to their practices, the Triple Bottom Line turns out to be the "good old-fashioned single bottom line plus vague commitments to social and environmental concerns" (Norman & MacDonald, 2004, p. 256; Smith & Sharicz, 2011). Thus, if the Triple Bottom Line framework is to be useful for this research, then it must be established in such a way that the resultant definition quantifies explicitly the commitments for each domain of sustainability, not just the economic domain.

Quantifying environmental sustainability. To quantify environmental sustainability, sustainability frameworks that prioritize environmental factors were considered. These included the Daly Rules (Smith, 2004a) and the strong sustainability framework (see Hediger, 2006) and are presented below.

The Daly Rules are a set of guidelines that define sustainable practices developed by Professor of Public Policy and economist Herman E. Daly (Smith, 2004a). The Daly Rules are considered because defining sustainable practices does provide an indirect definition or framework for sustainability. The Daly Rules are given in three parts:

(1) "renewable resources such as fish, soil, and groundwater must be used no faster than the rate at which they regenerate" (Smith, 2004a, p. 1);
(2) "nonrenewable resources such as minerals and fossil fuels must be used no faster than renewable substitutes for them can be put into place" (Smith, 2004a, p. 1);
(3) "pollution and wastes must be emitted no faster than natural systems can absorb them, recycle them, or render them harmless" (Smith, 2004a, p. 1).

These ideas have a large basis in thermodynamics, which means that quantifying sustainability through this conceptualization is tenable. It also relegates economic and social concerns to the aggregate of rates of consumption while focusing on the environmental aspect through rates of renewal. In terms of the Brundtland definition of sustainability, this directly connects the Daly Rules to the environmental domain of sustainability while indirectly connecting it to the economic and social domains.

Considering these rates of renewal and consumption further, Smith (2004a) explains the implications of such a model by summarizing the model in three different scenarios based on rates of consumption and rates of renewal (see Table 2.1). Thus, there are only two sustainable options according to Smith (2004a): steady-state sustainability and sustainable development. Moreover, metrics to measure these states for sustainability are available (see United Nations Division for Sustainable Development, 2001). Both steady-state sustainability and sustainable development directly connect with specific states of the environment here (Smith, 2004a).

Table 2.1
Definition of Sustainability Through Rates of Consumption and Renewal (Smith, 2004a)

Consumption of renewable resources	State of Environment	Sustainability
More than nature's ability to replenish	Environmental degradation	Not sustainable
Equal to nature's ability to replenish	Environmental equilibrium	Steady-state sustainability
Less than nature's ability to replenish	Environmental renewal	Sustainable development

The strong sustainability framework (see Hediger, 2006) is a framework that emphasizes environmental sustainability. Essentially, strong sustainability is:

> A physical principle which is founded upon the laws of thermodynamics and processes of biological growth. As a basic principle of resource management, it has a long tradition in forestry and has logically been extended to other domains of natural resource management. For instance, minimum criteria of "strong" sustainability are generally expressed in physical terms, saying that certain properties of the physical environment must be sustained. However, it is not clearly defined in the literature what it is that should be sustained. (Hediger, 2006, p. 362)

Essentially, this framework is saying that everything that occurs on earth is limited by the amount of incoming energy (i.e., the sun). That implies that the environmental domain of sustainability supersedes the other two domains because without the physical energy input that we receive from the environment, the social and economic domains of sustainability would not even exist. Also, the perspective of the strong sustainability framework parallels the Daly Rules as argued below.

Recall that the Daly Rules specified sustainability in physical terms (rates of consumption and renewal). The Daly Rules argue that sustainability of the economic and social domains should come in conjunction with environmental sustainability measured through rates of consumption and renewal (Smith, 2004a), while strong sustainability cannot specify sustainability of the economic and social domains because what is to be sustained is not defined in the literature (Hediger, 2006). In this sense, strong sustainability can be quantified through the Daly Rules. It may be that comparison of the rates of consumption and rates of renewal gives inherent rise to the simultaneous longevity of all three sustainability domains (economic, environmental, and social).

The Daly Rules offer the most accessible, specified, and quantifiable measures for the minimum level of performance for the Triple Bottom Line framework. Hence, for sustainability to be achieved in the environmental domain we must ensure that:

(1) "renewable resources such as fish, soil, and groundwater must be used no faster than the rate at which they regenerate" (Smith, 2004a, p. 1);
(2) "nonrenewable resources such as minerals and fossil fuels must be used no faster than renewable substitutes for them can be put into place" (Smith, 2004a, p. 1);
(3) and "pollution and wastes must be emitted no faster than natural systems can absorb them, recycle them, or render them harmless" (Smith, 2004a, p. 1).

Again, while the Triple Bottom Line framework (Kiewiet & Vos, 2007) calls for a minimum level of performance, that level of performance is not specified in the Triple Bottom Line framework. The steady-state sustainability (Smith, 2004a) criteria can satisfy this minimum level of performance with respect to the environmental domain. That is, environmental sustainability is defined as being met if the rates of resource consumption do not exceed rates of resource renewal.

Quantifying economic sustainability. Several perspectives of economic sustainability can be found across the literature. One such perspective is provided by Dyllick and Hockerts (2002), who argue that sustainability envisions a "more equitable and wealthy world in which the natural environment and our cultural achievements are preserved for generations to come" (p. 130). They continue that corporate sustainability was defined as "meeting the needs of a firm's direct and indirect stakeholders (e.g., shareholders, employees, clients, pressure groups, communities, etc.), without compromising its ability to meet the needs of future stakeholders as well" (Dyllick & Hockerts, 2002, p. 131). Dyllick and Hockerts (2002) posit that this definition is a revision of the Brundtland definition, choosing to focus on corporate interests (i.e., economic) instead of the generalized interests expressed in the Brundtland definition. However, envisioning a world that preserves the natural world (i.e., environmental sustainability) and cultural achievements (i.e. social sustainability) necessitates that at least environmental and social factors are indirectly addressed or focused upon.

Two similar frameworks to that of Dyllick and Hockerts (2002) can be identified in the very weak sustainability framework and the weak sustainability framework (see Hediger, 2006). The very weak sustainability framework contends that there must be constant consumption per capita and requires that the overall production capacity (i.e., the economic capacity) be maintained (Solow, 1974, 1986). In contrast, weak sustainability "requires that the welfare potential of the overall capital base remains intact" (Hediger, 2006, p. 362). In essence, weak sustainability is a step above very weak sustainability because it allows for the value of non-consumptive items (such as the environment, wildlife, etc.) to be included in the value of the overall capital. However, quantifying these constructs is vague in this framework; what non-consumptive items should be considered as part of the overall capital? While the economic domain of

sustainability is easily quantified in terms of a capital base (i.e., worth in dollars), environmental and social value can sometimes be measured (see United Nations Division for Sustainable Development, 2001). However, how are these equated to the overall capital base? Both the very weak and weak sustainability frameworks leave this question unanswered, though the latter arguably better quantifies the economic domain of sustainability.

Arrow, Dasgupta, Goulder, Mumford, & Oleson (2012) offer a more comprehensive framework of economic sustainability that integrates and quantifies the inherent wealth of natural capital (i.e., the environmental domain) and health capital (i.e., an aspect of the social domain). This view addresses the shortcomings of the very weak and weak sustainability frameworks. However, in the context of this study, the researcher perceived that post-secondary STEM students would not generally be able to conceive of the economic domain of sustainability as incorporating such complex measures of natural and health capital. For the purposes of this study, despite the greater comprehensiveness Arrow et al.'s (2012) framework offers in the economic domain of sustainability, weak sustainability was thus used to define the economic domain sustainable state. That is, economic sustainability is defined as the state when generalized production capacity of an economy is maintained intact, such as to enable at least constant consumption per capita through time (Solow, 1974, 1986).

Quantifying social sustainability. Frameworks that offer definitions of sustainable states for the social domain of sustainability seemed elusive in the literature. With respect to the social domain, Gagnon, Leduc, and Savard (2009) offer an indirectly related sustainability framework, contending that a development is sustainable if “it allows every people [sic] globally to at least meet their basic needs, if it provides individuals in a given society equal opportunities to increase their quality of life, and if it provides future generations increasing opportunities” (p. 1467). It should be noted that Gagnon et al. (2009) frame their definition within the context of an interdisciplinary group project setting where engineers work with members of other disciplines to fulfill the needs of clients. While the work of Gagnon et al. (2009) focuses on economic, environmental, and social factors, their definition focuses strongly on sustaining social equity. Along similar lines, Seliger, Khraisheh, and Jawahir (2011) offer that sustainable

manufacturing is dedicated to sustainable products and processes, and that these "conserve energy and natural resources, have minimal impact upon the natural environment and society, and adhere to the core principle of considering the needs of the present without compromising the ability of future generations to meet their own needs" (p. v).

Another sustainability framework, the Natural Step framework (Nattrass & Altomare, 1999), contends that:

> In the sustainable society, nature is not subject to systematically increasing (1) concentrations of substances extracted from the Earth's crust; (2) concentrations of substances produced by society; or (3) degradation by physical means; and in that society (4) human needs are met worldwide. (p. 23)

Environmental sustainability is directly connected to the first three elements of the Natural Step framework, while the final element directly connects to social sustainability. "Human needs" can be interpreted in many ways, and this framework does not specify quantifiable constructs to assess social sustainability in this context. For the purposes of this study, human needs are defined as clean water for consumption and use, food, shelter, clothing, healthcare, and education. Since cultures and peoples vary greatly in terms of how they live, it would be outside reasonable expectation that anything beyond this requirement be considered in the social domain of sustainability on a global scale. While this may be the only global directive within the social domain of sustainability, there are other social aspects to consider at non-global levels. Moreover, while education may not generally be considered a human need, it will have to be considered as such for sustainability. The argument here is simple; if global population levels continue to increase, they will eventually outgrow the limits of the planet to support us (see Attenborough, 2011; Franck, von Bloh, Müller, Bondeau, & Sakschewski, 2011). We cannot ever hope to be sustainable in an unbounded growth scenario since our physical environment naturally limits our population. Hence, education must be employed to insure that all global communities understand this.

The quantifiable definition of sustainability. The synthesis of the quantifying economic, environmental, and social sustainability sections led to the creation of the *quantifiable definition of sustainability*. The quantifiable definition of sustainability is labeled as "quantifiable" since the definition specifies measureable variables or goals to

assess whether a development is sustainable or not. The *quantifiable definition of sustainability* is as follows:

> A development is sustainable if it meets the minimum level of performances in each of the areas of concern:
> (1) Environmental Domain:
> a. renewable resources such as fish, soil, and groundwater must be used no faster than the rate at which they regenerate (Smith, 2004a, p. 1)
> b. nonrenewable resources such as minerals and fossil fuels must be used no faster than renewable substitutes for them can be put into place (Smith, 2004a, p. 1)
> c. pollution and wastes must be emitted no faster than natural systems can absorb them, recycle them, or render them harmless (Smith, 2004a, p. 1)
> (2) Economic Domain: that the generalized production capacity of an economy is maintained intact, such as to enable at least constant consumption per capita through time (Solow, 1974, 1986)
> (3) Social Domain: at least basic human needs (clean water for consumption and use, food, shelter, clothing, healthcare, and education) are met worldwide (Nattrass & Altomare, 1999, p. 23)

While this definition of sustainability makes each domain of sustainability quantifiable, it cannot be expected to be representative of sustainability at all levels. This reflects the sentiment that situational additions for sustainability must be considered (Kiewiet & Vos, 2007; Shriberg, 2002).

Care must be given not to interpret the dimensions of the quantifiable definition of sustainability as unrelated or orthogonal to one another. The measured interdependence of these domains of sustainability has not been studied to the researcher's knowledge. Indeed, part of the mission of sustainability is to understand the interdependence and independence of these dimensions with one another as well as how they collectively form the concept of sustainability.

Limitations of the quantifiable definition of sustainability. The presented collection of sustainability definitions and frameworks offer a diverse look at this expansive topic. While the Brundtland definition of sustainability may be dominant in the literature and there are some sustainability frameworks to help guide us towards sustainability, it still leaves room for disagreement in the field. However, though a continuing discussion on the definition of sustainability should be pursued in the field, Trzyna (1995) offers that "sustainability is not a precise goal but a criterion for attitudes

and practices" (p. 16). Munro (1995) echoed this idea, claiming that sustainability is a "continuous and iterative process, through and throughout which experience in managing complex systems is accumulated, assessed, and applied" (p. 34). Bell and Morse (2008) offer that "the very holistic and anthropocentric essence of sustainability continues to elude attempts at objective analysis and assessment" (p. xvii). This thought has been prevalent in the literature (see Gibbon, Lake, & Stocking, 1996; Izac & Swift, 1994; Kidd, 1992; Robinson, 2004; Schaller, 1989). Some have even concluded that this must mean that sustainability is a futile effort (Morris, 2012). However, our greatest milestones as a species have often come at the precedent of great and overwhelming odds against us. The quantifiable definition of sustainability provides such a perspective for the direction of the STEMSEI.

As one last consideration, it has been argued in business that sustainability will weaken the economy by placing such restrictions on how we interact with the environment (Ameer & Othman, 2012; Anderson, 2009). This could be interpreted to mean that the environmental and economic portions of the quantifiable definition of sustainability are contradictory. However, there is growing literature to show sustainability can be profitable (Ameer & Othman, 2012; Anderson, 2009). To further reconcile any divide between economic and environmental sustainability, we must change the way we do business, as advocated for in various ways across the literature (see Ameer & Othman, 2012; Anderson, 2009; Dyllick & Hockerts, 2002; Fairbrass & Zueva-Owens, 2011; Johnston, Everard, Santillo, & Robèrt, 2007; Norman & MacDonald, 2004).

In short, given the complexities under and in which sustainability operates, the quantifiable definition of sustainability cannot be considered static; it will likely require revision in the future to encompass new knowledge of sustainability and possibly to accommodate changes in the perception of sustainability by post-secondary STEM students. This necessarily means that the STEMSEI may not function as intended in future populations due to changes in sustainability or the perception of it.

Theoretical framework overview. Concisely, this study views sustainability engagement along two dimensions. First, the engagement portion of the construct has three subdomains: emotional, cognitive, and behavioral. Second, sustainability itself also

has three subdomains: economic, environmental, and social. Effectively, means there are 9 subdomains of sustainability engagement (see Table 2.2). While this delineates sustainability engagement into many smaller pieces, and thus, can be perceived as too complicated, such a scope on sustainability engagement is needed to avoid inadequate preoperational explication of the construct. Inadequate preoperational explication of a construct occurs when specific facets of a construct are unaccounted for, which gives an incomplete view of the construct (McMillan & Schumacher, 2010). That incomplete view of the construct can also introduce bias and error in measurements. As such, a complete, encompassing view of sustainability engagement must include all 9 of these aspects of sustainability engagement.

Table 2.2
Delineations of Sustainability Engagement as Proposed by the STEMSEI

		Sustainability Domains		
		Economic	Environmental	Social
Engagement Domains	Emotional	1	4	7
	Cognitive	2	5	8
	Behavioral	3	6	9

Expected Outcomes Across the Literature in Sustainability Related Measures

This brief section highlights some of the results in sustainability related measures that could inform STEMSEI development. Only results relating to the environmental domain of sustainability were found.

First, Johnson, Bowker, and Cordell (2004) conducted a study comparing ethnic differences in environmental belief and behavior. For their study, Johnson et al. (2004) used the National Survey on Recreation and the Environment (see United States Department of Agriculture, n.d.) and the New Ecological Paradigm (see Dunlap et al., 2000) as quantitative measures of their constructs. Given that Johnson et al. (2004) focused their study only on environmental belief and behavior (i.e., not economic or social issues), this use of the New Ecological Paradigm was seen as appropriate. Johnson et al.'s (2004) results identified statistically significant differences in both environmental belief and behaviors due to race. Specifically, Caucasians had higher measures of pro-

environmental beliefs when compared to African-Americans and foreign-born Latinos. Considering this, there may be statistically significant differences in scores generated by the environmental scale of the STEMSEI when comparing race.

Olli, Grendstad, and Wollebaek (2001) used a battery of survey instruments to identify different trends in environmental beliefs based on social context. Olli et al. (2001) found that females were more engaged in environmental issues but less engaged in environmental activism. Johnson et al. (2004) also found similar statistically significant differences between males and females in environmental belief and behavior. With respect to the STEMSEI, the context of engagement in these studies could be either emotional, cognitive, or behavioral. Hence, statistically significant differences between males and females were considered across the STEMSEI with no clear expectation of how differences would manifest across emotional, cognitive, and behavioral engagement.

Third, Lalonde and Jackson (2002) argued that changes to the New Environmental Paradigm (see Dunlap & Van Liere, 1978) should be considered due to the "Judeo-Christian views of unlimited natural abundance" (p. 33). This may mean that Judeo-Christians would have lower scores on the environmental scale due to this world view. Thapa (1999) makes similar arguments, suggesting that the "human exemptionalist" paradigm that is rooted in Judeo-Christian thoughts promotes the view that humans can overcome all problems with human ingenuity alone. Based on these works, statistically significant differences in STEMSEI environmental measures might be expected along past and/or present religious affiliation.

From all of these results, it was hypothesized that there would be statistically significant differences in environmental engagement measures based on race, gender, and religious beliefs.

Copyright © David L. Little II 2014

Chapter III

Methodology

The purpose of this three phase multi-methods study was to create, evaluate, and validate the STEM Sustainability Engagement Instrument (STEMSEI), which measures sustainability engagement in post-secondary science, technology, engineering, and mathematics (STEM) majors. Specifically, the following research questions were explored and corresponding hypotheses were tested:

1. To what extent can a convergent theoretical framework for sustainability engagement in post-secondary STEM students be achieved between sustainability experts across the STEM disciplines?
   Hypothesis: A convergent theoretical framework for the STEMSEI is possible.
2. To what extent can items that measure unique types of sustainability engagement and can be interpreted by post-secondary sustainability educators and post-secondary STEM students be developed across the STEM disciplines?
   Hypothesis: The factor structure of responses to the STEMSEI will distinguish among either sustainability domain (see Kiewiet & Vos, 2007) or engagement type (see Fredricks, Blumenfeld, & Paris, 2004).
3. To what extent does the STEMSEI produce interpretable and useable/meaningful results with respect to sustainability education at the post-secondary level?
   Hypothesis 1: There will be statistically significant differences in environmental engagement measures based on race (see Johnson, Bowker, and Cordell, 2004), gender (see Johnson, Bowker, and Cordell, 2004; Olli, Grendstad, & Wollebaek, 2001), and religious beliefs (Lalonde & Jackson, 2002; Thapa, 1999).
   Hypothesis 2: There will be statistically significant differences in sustainability engagement (as indicated by scores produced by the STEMSEI) across STEM majors.
   Hypothesis 3: There will not be statistically significant differences in sustainability engagement (as indicated by scores produced by the STEMSEI) across classification (undergraduate/graduate student).

Though behavioral engagement is a core component of the sustainability engagement framework, the researcher did not have the time or resources to develop this portion of the STEMSEI as well. This decision was also influenced by the fact that it was not known if a quantitative instrument could be developed to measure sustainability engagement in post-secondary STEM students. Moreover, since measuring behaviors via self-report methods can lead to inaccurate measures and invalid instruments (see Fan et

al., 2006), the researcher thought development of successful measures of emotional and cognitive engagement were most probable. As such, behavioral engagement is not considered further in the methodology. This is a limitation of the STEMSEI as it is presented here. Measures of behavioral engagement in sustainability should be developed in the future.

Instrument development is an iterative process that involves qualitative and quantitative methods of quality assessment (Benson & Clark, 1982; Downing, 2006; Linn, 2006). Due to this, different study designs were employed across the three phases of this study. The three phases of this study were developed from a four-phase methodology for instrument development in occupational therapy recommended by Benson and Clark (1982). Despite the methodology being over three decades old, it is still recommended for instrument development in occupational therapy (Wæhrens, 2010) and any area developing an instrument (Downing, 2006; Linn, 2006).

Phase 1 of this study focused on instrument planning (Benson & Clark, 1982), where the purpose, constructs of interest, and objectives of the STEMSEI were developed. Phase 1 (instrument planning) encompassed a review of the literature (see Chapter II) and a qualitative study using content expert input and feedback to validate the theoretical framework used to develop the STEMSEI (see Chapter II).

Phase 2 focused on instrument construction, which included instrument specifications and developing an item bank (Benson & Clark, 1982) for the STEMSEI. Phase 2 (instrument construction) was a concurrent mixed methods study to guide instrument and item development using input and feedback from content experts and post-secondary STEM students.

Finally, phase 3 of this study involved the quantitative and qualitative evaluation of the instrument (Benson & Clark, 1982). Phase 3 (instrument evaluation) was a sequential mixed methods study that evaluated the functioning of the STEMSEI in post-secondary STEM students.

Benson and Clark (1982) also recommend a fourth phase for instrument development that was not explored fully in this study: the validation phase. In the validation phase of instrument development, different researchers, over the course of several years, test the validity of the instrument in question by determining the extent to

which an instrument accurately and reliably measures the construct of interest (Benson & Clark, 1982). This final recommended phase could not be completed in this study due to limitations of time and resources. Moreover, the validation phase should be completed with multiple independent researchers' input.

Though the validation phase from Benson and Clark's (1982) methodology was not completed, it was essential to gather validity evidence supporting the STEMSEI throughout this study. Such validity evidence would ensure the methods and results of each phase were based on evidence supporting the structure, use, and methodology employed in the STEMSEI's development. Messick (1990) argued that validity was a multi-faceted construct and that validity is not simply "present" or "not present"; rather, validity exists as a gradient or there are different degrees of validity. Using Messick's (1990) progressive validity matrix, validity assessments were developed for each phase of the development of the STEMSEI. These validity assessments focused on construct validity, value implications, and relevance/utility (Messick, 1990). Social consequences (Messick, 1990) were not examined in-depth since the consequential use of the STEMSEI cannot be fully predicted a priori. With respect to intended use (a tool to evaluate sustainability engagement in post-secondary STEM students), hypothesized social consequences include (1) improvements to sustainability curricula to promote sustainability engagement and/or (2) increased or decreased perceived importance of sustainability engagement by social and political institutions. However, since this study concerns the development of the STEMSEI and not its actual use, social consequences of the STEMSEI are not examined beyond these speculations. On the other hand, construct validity, value implications, and the relevance/utility of the STEMSEI was assessed throughout the study.

The results of this study were meant to (1) provide post-secondary sustainability educators an instrument appropriate to assess sustainability engagement in post-secondary STEM students, (2) to provide means to establish norms for sustainability engagement for reference and research, and (3) to inform the development of future sustainability education instruments that may measure similar or related constructs of interest.

Phase 1: Instrument Planning

Benson and Clark (1982) recommend that the first phase of instrument development focus on instrument planning. There are three main steps in this phase (see Figure 3.1; developed from Benson & Clark, 1982). These steps develop the instrument purpose, target population, constructs of interest, objectives, and test format (Benson & Clark, 1982). The following section details the study design for this phase, including the sample and data collection and analysis techniques where applicable.

Phase 1, step 1: Development of purpose, target population, and domains. The purpose of the STEMSEI was to measure sustainability engagement across post-secondary STEM majors. Measurement was framed from the perspectives of emotional, cognitive, and behavioral engagement across the economic, environmental, and social domains of sustainability.

With respect to the target population, having an instrument that measures sustainability engagement across a wide age group would be beneficial. However, an instrument measuring sustainability engagement across students at the elementary, secondary, and post-secondary education was not seen as tenable due to cognitive differences in these populations. Due to this, the target population was first restricted to only post-secondary students. The target population was then further restricted to post-secondary STEM majors to (1) help improve generalizability of any findings, (2) simplify

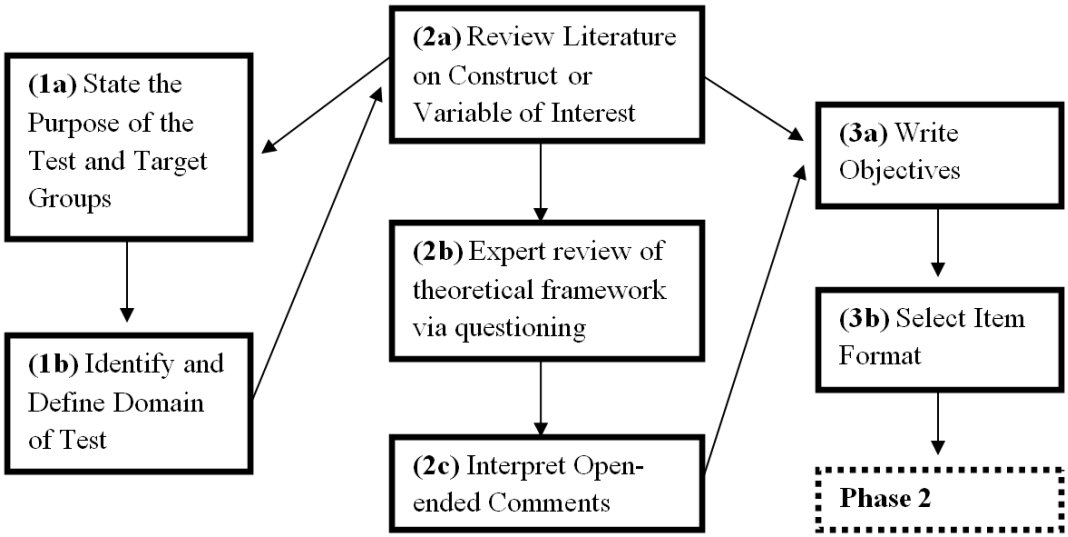

Figure 3.1. Study design of phase 1, instrument planning. Individual columns represent different steps in this phase. Phase 2 begins after step 3b is completed.

instrument development by limiting the content focus of items, and (3) because STEM education was the specialization of the researcher.

To the researcher's knowledge, sustainability engagement as a construct lacked both a theoretical structure and a formal definition in the literature. This meant that no clear domains were evident. The literature did suggest, however, that sustainability engagement was not a unidimensional construct (i.e., knowledge for sustainability versus action for sustainability; see Christensen et al., 2009; Davis et al., 2003; Hodson, 2003; O'Connell et al., 2005; Thapa, 1999; Van Kerkhoff & Lebel, 2006; Wright et al., 2009). Both knowledge and action for sustainability were hypothesized to relate to elements of education pedagogy and behavior change (Frisk & Larson, 2011). A similar framework for school engagement (Fredricks, Blumenfeld, & Paris, 2004) encompassed elements of both education pedagogy and behavior change. Fredricks et al. (2004) theorized school engagement was composed of three sub-domains: emotional, cognitive, and behavioral. Hence, the construct of interest for the STEMSEI (sustainability engagement) was delineated simultaneously along the domains of sustainability (economic, environmental, and social) and the domains of engagement (emotional, cognitive, and behavioral).

Phase 1, step 2: Literature review and expert panel review. The literature review in step 2a (see Figure 3.1) of Benson and Clark's (1982) methodology was completed and is presented in Chapter II. From that review of the literature, the quantifiable definition of sustainability and the sustainability engagement framework (see Chapter II) were theorized. Validation of the quantifiable definition of sustainability and the sustainability engagement framework was important for two reasons. First, the researcher wanted to ensure that the quantifiable definition of sustainability and the sustainability engagement framework were relevant, interpretable, and useful for post-secondary sustainability education researchers. This would ensure the relevance/utility (Messick, 1990) of the STEMSEI with intended users (sustainability education researchers). This would also promote ease of future use of the STEMSEI with sustainability education researchers. Second, content expert evaluation of the theoretical framework would help uncover any gaps or inadequacies in the theories for measuring sustainability engagement. This would help ensure construct validity (see Messick,

1990). Gaps or inadequacies in fully defining and explaining the construct of interest would introduce error in any measures of the constructs due to inadequate preoperational explication of the constructs (McMillan & Schumacher, 2010). Inadequate preoperational explication of the constructs is when a construct conceptualization or framework omits a facet of the construct (i.e., a subdomain) that results in an incomplete or inexplicable operational conceptualization of the construct. For example, if the STEMSEI only measured environmental engagement, it would not be an accurate and representative measure of sustainability engagement as was argued in Chapter II with respect to the New Ecological Paradigm (Dunlap, Van Liere, Mertig, & Jones, 2000). This is because sustainability consists of more than environmental issues (Lemonick, 2009; McDonough & Braungart, 2002), and includes social and economic issues as well (Bencze, Sperling, & Carter, 2011; Erdogan, 2010; Sekulic, 2011; Wainwright, 2010). Hence, the economic and social aspects of sustainability must be accounted for as well. Identifying further aspects of sustainability that should be measured was one question to be resolved with the content experts.

Steps 2b and 2c (see Figure 3.1) pertain to the validation of the comprehensiveness and preoperational explication of the constructs as operationalized through the sustainability engagement framework and the quantifiable definition of sustainability. This validity assessment was completed using data gathered from semi-structured interviews (see Appendix B) that were conducted using the Delphi method (see Okoli & Pawlowski, 2003) with content experts in sustainability education. In the Delphi method, separate interviews using the same interview protocol are conducted with multiple participants across multiple rounds (Okoli & Pawlowski, 2003). The Delphi method (see Okoli & Pawlowski, 2003) was chosen because it is commonly used in framework development due to the method's ability to produce convergent frameworks (Okoli & Pawlowski, 2003). Consequently for this study in using the Delphi method (see Okoli & Pawlowski, 2003), both the quantifiable definition of sustainability and the sustainability engagement framework would be refined from multiple disciplinary perspectives given the diverse backgrounds of the content experts in the sample. Refining these two ideas from multiple perspectives across the STEM fields would allow

for the (1) ease of use of the STEMSEI across the STEM fields (see Appendix A) and (2) ease of dissemination of results from the STEMSEI. Moreover, as argued by Benson and Clark) 1982), this expert review would offer content, construct, and domain evidence (see Messick, 1990) as well as evidence for content relevance and representativeness (see Messick, 1990). First, this method would produce a definition of sustainability and a framework for sustainability engagement that was convergent across multiple STEM disciplines, as it is commonly used in framework development (Okoli & Pawlowski, 2003). Second, both the quantifiable definition of sustainability and the sustainability engagement framework would then be refined from multiple disciplinary perspectives to allow for the diverse use and dissemination of the STEMSEI. While the semi-structured interview protocol (see Appendix B) used in the Delphi method (see Okoli & Pawlowski, 2003) served to produce a convergent, accessible, and comprehensive definition of sustainability and sustainability engagement framework, this phase also developed a foundation on which to assess the STEMSEI's validity in later phases.

The semi-structured interview protocol (see Appendix B) for the Delphi method was designed to delineate and refine construct definitions and to identify other possible sub-domains. This would establish construct and domain evidence (see Messick, 1990) for the STEMSEI as well as uncover any inadequate preoperational explication of the constructs. The utility and usefulness of measuring the construct of interest (sustainability engagement) was also verified in the interviews, establishing relevance/utility (see Messick, 1990) of the STEMSEI. Finally, the interviews established baselines for certain methodological questions, such as the importance of the domains of sustainability (economic, environmental, and social) and of engagement (emotional and cognitive). In turn, this established content relevance and representativeness (see Messick, 1990) of the STEMSEI.

Sample. For step 2 of phase 1, a convenience sample of six content experts in sustainability education were identified from one post-secondary institution in the mid-South United States. The researcher identified only six content experts to participate for two reasons. First, due to limited time of the content experts and the researcher, a smaller sample was necessary. Second, the nature of the semi-structured interviews for this portion of the study asked yes-no questions with explanations provided by the content

experts as needed. If answers across the initial sample were unanimous or all but one content expert agreed, this was interpreted to mean that content experts beyond the initial 6 were not necessary. This follows recommendations for qualitative studies as outlined by McMillan and Schumacher (2010, p. 328).

The content experts were distributed so that equal representation across the domains of sustainability (economic, environmental, and social) was possible, with some experts providing feedback on multiple domains of sustainability. Equal representation across all domains of sustainability was seen as necessary to avoid tendencies towards any one domain or specifically overlooking other domains.

Three experts were engineering professors with varied research interests in sustainability education. All three were involved in research concerning environmental issues, including sustainable manufacturing, sustainable supply chains, and minimizing environmental impact from manufacturing. Two of the engineering professors also studied social sustainability issues, including equitable treatment of workers and equitable distribution of goods and resources. The third engineering professor studied economic sustainability issues, such as maximizing profit of sustainable designs.

Two other experts were business professors with research interests in sustainable business practices and economic impacts of social sustainability initiatives. These experts contributed to the economic and social domains of sustainability.

The final content expert was a literacy education professor with no expertise within STEM. This content expert did have research interests in sociological issues concerning sustainability, such as equity and discourse analysis. Given this content expert's specialization in literacy and discourse analysis, this content expert was seen as uniquely qualified of all six content experts to review item phrasing and terminology in the STEMSEI. Hence, the lack of expertise in STEM was not seen as a hindrance compared to the benefits gained in item review.

Collection and analysis of data. Steps 2b and 2c of Benson and Clark's (1982) methodology (see Figure 3.1) pertain to, respectively, content expert review of the theoretical framework of the instrument in question and interpretation of content expert feedback. A qualitative study design using the Delphi method (see Okoli & Pawlowski,

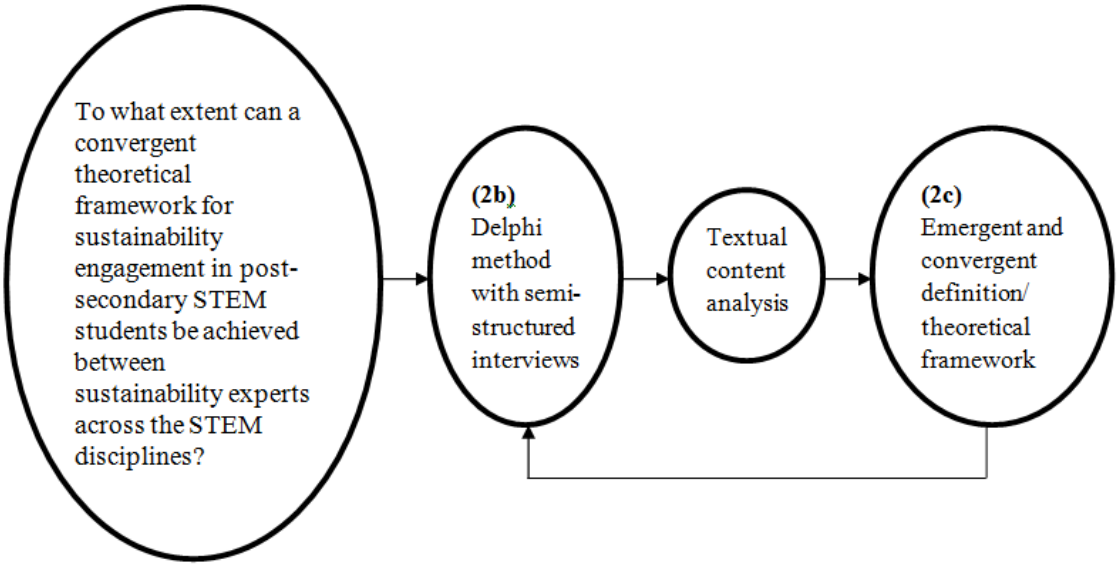

Figure 3.2. Qualitative study design for steps 2b and 2c of phase 1. Recall steps 2b and 2c were, respectively, expert review of theoretical framework and interpretation of responses.

2003) was developed (see Figure 3.2) for these two steps. A semi-structured interview protocol (see Appendix B) was used to gather open-ended data from the content experts. The semi-structured interview protocol (see Appendix B) was developed to assess the completeness and relevance of the quantifiable definition of sustainability and the sustainability engagement framework.

In the Delphi method, multiple rounds of interviews are conducted with each participant separately (Okoli & Pawlowski, 2003). Only two rounds were necessary in this study. All interviews were conducted by the researcher. During interviews, the researcher took notes on responses and authenticated data using member check (i.e., narrative accuracy checks) (see Morse, Barrett, Mayan, Olson, & Spiers, 2002). Interviews were not recorded. Data was authenticated with narrative accuracy checks at the end of each question by first rereading the notes taken by the researcher. Then, participants were asked if the narrative was accurate or if alterations should be made. After each round, results of the interviews were interpreted and coded using textual content analysis (see Carley, 1993). In textual content analysis, for a given data set, codes are developed that identify common themes across the data (Carley, 1993). Qualitative data (i.e., notes taken during the interviews) can then be converted into quantitative data (i.e., frequencies) by matching qualitative data to the developed

Table 3.1
Content Analysis Codes for Content Expert Interviews

Code	General Content
Relevance/Utility	Feedback relates to relevance/utility of quantifiable definition of sustainability and/or the sustainability engagement framework (i.e., if measures of sustainability engagement are important)
Sustainability Domains/Constructs	Feedback relates to the domains of sustainability as presented in the quantifiable definition of sustainability (i.e., if the economic, environmental, and social domains are sufficient)
Engagement Domains/Constructs	Feedback relates to the domains of sustainability engagement as presented in the sustainability engagement framework (i.e., if the emotional, cognitive, and behavioral domains are sufficient)
Methodology Feedback	Feedback relates to the methodology employed in creating a measure of sustainability engagement (i.e., equal representation across domains)

codes for the textual content analysis. The codes used in this study for the textual content analysis are presented in Table 3.1. After the data were coded, the data were synthesized by identifying patterns (McMillan & Schumacher, 2010). These patterns were interpreted as suggestions by the content experts which were then used to edit and modify the quantifiable definition of sustainability and the sustainability engagement framework. After edits to the quantifiable definition of sustainability and the sustainability engagement framework were made, a subsequent round began, repeating the same process. This process was repeated until there were no additional recommendations from the content experts, which took only two rounds in this study.

Phase 1, step 3: Objectives and item format. Following step 2 of phase 1, the quantifiable definition of sustainability and the sustainability engagement framework were determined to be convergent, interpretable, and comprehensive (i.e., no changes were made). This definition of sustainability and framework for sustainability engagement then provided a foundation for further developing the STEMSEI. To

Table 3.2
Developed Objectives for the STEMSEI

Objective	Description
1 - Sustainability Domains	This instrument will assess engagement across all domains of sustainability (economic, environmental, and social). See the quantifiable definition of sustainability for definitions of each domain of sustainability.
2 - Engagement Domains	This instrument will assess emotional, cognitive, and behavioral engagement with respect to sustainability. See the sustainability engagement framework for definitions of each engagement type.
3 - Differentiation	When differences exist in the population that are supported in the literature, this instrument will differentiate in the population accordingly.

complete step 3a of phase 1 (see Figure 3.1), instrument objectives were then developed (see Table 3.2). Objectives 1 and 2 (see Table 3.2) ensured complete content representation across both sustainability (economic, environmental, and social) and engagement types (emotional, cognitive, and behavioral). This ensured the STEMSEI would have adequate preoperational explication of the construct (i.e., sustainability engagement). Objective 3 (see Table 3.2) ensured that the STEMSEI replicated results that were already predicted in prior literature/research.

To complete step 3b in phase 1 (see Figure 3.1), an item response format needed to be selected. There are several response scales available for instruments (see Smith, 2004b). First, a polytomous scale was seen as preferable to a dichotomous scale for the STEMSEI. Polytomous scale options are more precise than dichotomous scale options with respect to measurement (Smith, 2004b). This is because polytomous scales measure direction and extremity where-as dichotomous scales only measure direction (Smith, 2004b). Hence, to promote variance between responses and, thus, accuracy of discrimination between respondents, a polytomous scale option was chosen for the STEMSEI.

Both bipolar and unipolar scale options were also included. Bipolar scale options are typically used to represent opposite valence in equivalent intensity (i.e., strongly disagree, disagree, agree, strongly agree) (Alwin, 2010). On the other hand, unipolar scale options typically differ in intensity or frequency but not valence (i.e., to no extent, to a small extent, to some extent, to a great extent) (Alwin, 2010). According to Alwin (2010), bipolar response scales are typically used for affective items (i.e., emotional engagement) while unipolar response scales typically measure frequencies (i.e., cognitive engagement). However, since emotional sustainability engagement relates to opinions or preferences towards sustainability (see Chapter II), unipolar response scales may be appropriate for some emotional engagement items depending on what the item measures. For example, items measuring concern for certain sustainability topics should utilize a unipolar response scale since an individual either has concern for a topic (positive latency) or no concern for a topic (neutral latency). Negative latency is not possible in this situation, and hence a bipolar scale is not possible with respect to items related to concern.

Considering all of this, unipolar and bipolar response scales were selected for the emotional engagement items while the cognitive engagement items used only a unipolar response scale. Selection of response scales for the instrument then led to the next phase of the study - instrument construction.

Phase 2: Instrument Construction

After the conclusion of phase 1, a convergent theoretical framework (see Chapter II) was identified and validity evidence had been attained. These frameworks provided a foundation to build the STEMSEI upon, and items were developed that measured sustainability engagement in the population while reflecting the nuances of the theoretical framework. This transitioned the study to phase 2, the instrument construction phase (see Figure 3.3).

Phase 2, step 4: Instrument blueprint and item writing. Step 4a (develop instrument blueprint) (see Figure 3.3) was completed using content expert feedback gathered from the semi-structured interviews (see Appendix B) in the previous phase. With that feedback, the importance of the domains of sustainability (economic, environmental, and social) with respect to one another was established. Specifically, the

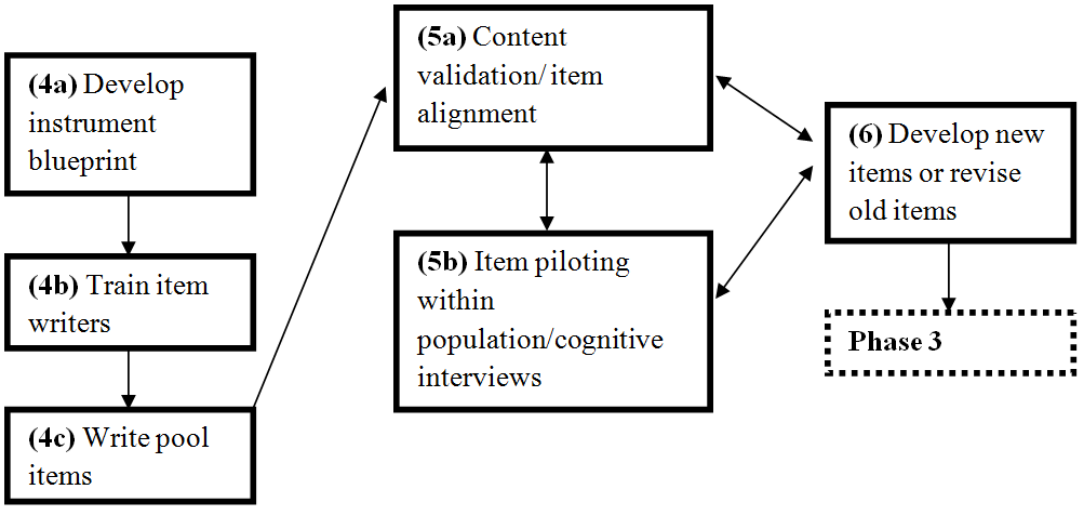

Figure 3.3. Study design of phase 2, instrument construction. Individual columns represent different steps in this phase. Phase 3 begins after step 6 is completed. *Note.* The names of steps 5a and 5b were modified from the language used by Benson and Clark (1982). These changes were made to allow the names of the steps to coincide with language used in more recent methodologies corresponding to these steps.

content experts were asked in the semi-structured interview in Appendix B if each domain of sustainability should be equally represented or if one was more important than another. A similar question was asked regarding the engagement types (emotional and cognitive). Using content expert feedback, instrument blueprint guidelines for the ratio of items representing each domain of sustainability (economic, environmental, and social) and engagement type (emotional and cognitive) were established (see Table 3.3); for details, see Chapter IV.

Table 3.3
STEMSEI Blueprint

		Emotional Engagement	Cognitive Engagement	Totals
Sustainability Domains	Environmental	$16.\bar{6}\%$	$16.\bar{6}\%$	$33.\bar{3}\%$
	Economic	$16.\bar{6}\%$	$16.\bar{6}\%$	$33.\bar{3}\%$
	Social	$16.\bar{6}\%$	$16.\bar{6}\%$	$33.\bar{3}\%$
Totals		50.0%	50.0%	100.0%

In Step 4b of phase 2 (train item writers) (see Figure 3.3), Benson and Clark (1982) assume that multiple content experts will be available to write items. However, the researcher was the only item writer for this study given that there were no resources to employ outside item writers. Though Benson and Clark (1982) do not formally define the activities of this stage, the researcher took it as implied that item writers must (1) fundamentally understand the construct of interest, (2) understand the instrument blueprint, (3) understand the instrument objectives, and (4) are familiar with at least the basics of optimal item writing practices. The first three elements of item writer training were completed by the researcher by being the facilitator of this study. Since the researcher developed the instrument and its objectives with feedback from content experts, this was seen as sufficient for the first three elements of item writer training. For the final element of item writer training, the researcher reviewed literature pertaining to item writing, including Krosnick and Presser's (2010, p. 264) guidelines for item writing, took a course in Measurement Theory and Techniques that focused on item writing practices, and was a member of the item writing team for a multi-million dollar National Science Foundation Discovery Research K-12 project that utilized a test writing company for its training, writing, and validation. These included recommendations such as using simple wording for items, not using ambiguous words, avoiding single or double negations, and avoiding items pertaining to more than one thing (i.e., double-barreled questions) (Krosnick & Presser, 2010, p. 264). Other recommendations for question ordering included presenting items of similar content together as well as placing "easier" to endorse items first (Krosnick & Presser, 2010, p. 264).

While constantly referring back to the instrument blueprint, instrument objectives, and following the item writing recommendations from both previous experience and from Krosnick and Presser (2010), the researcher then created a bank of items. The developed items tapped the construct of interest (sustainability engagement) along the different domains of sustainability (economic, environmental, and social) and two of the different types of engagement (emotional and cognitive). Since the STEMSEI was designed to measure sustainability engagement across all STEM majors, care was given to write items that were endorsable for all STEM majors and across the varying degrees of

endorsability. This was to ensure content relevance and representativeness while also maintaining the relevance/utility element of validity (see Messick, 1990).

Phase 2, step 5 and step 6: Content validation and item piloting. Steps 5 and 6 of phase 2 pertained to content validation/item alignment, item piloting/cognitive interviews, and item revision/additions. In step 5a, content validation/item alignment was established by having content experts judge in which cell of the instrument blueprint an item fits or aligns to (Benson & Clark, 1982). The content experts evaluated items and decided if it measures economic, environmental, and/or social sustainability engagement as well the type of engagement (emotional, cognitive, and/or behavioral). Items that fit in multiple cells of the instrument blueprint (i.e., both the "emotional economic engagement" cell and the "emotional environmental engagement" cell) were considered for revision or deletion. In step 5b, item piloting involved having members of the instrument's target population (i.e., post-secondary STEM majors) judge (1) clarity of directions, (2) item wording clarity, and (3) make recommendations to improve clarity (Benson & Clark, 1982). Finally in step 6, old items were revised based on prior feedback or new items were written (Benson & Clark, 1982).

No concrete methodology for assessing steps 5a or 5b were given in Benson and Clark's (1982) work. Hence, methodologies for these two steps were developed from current research methodologies that accurately reflected Benson and Clark's (1982) intent of these steps. Since any item revisions or item additions in step 6 might require subsequent rounds of steps 5a and 5b, these steps were implemented concurrently using a concurrent mixed methods design (see Creswell, 2002) which is presented in the following subsections.

Sample. First, the same six content experts from phase 1 participated in step 5a (i.e., content validation/item alignment). Webb (1997) recommends at least five content experts for alignment tasks based on tests developed for assessment purposes in P-12 education, so six content experts were seen as sufficient.

Additionally, 12 post-secondary STEM students participated in cognitive interviews. In cognitive interviews, interviewees are asked to read a sample item and explain their thoughts about the item and their thought process in determining their response to the item (Baker, Crawford, & Swinehart, 2004; Willis, 1999). The sample of

12 interviewees followed recommendations from Baker, Crawford, and Swinehart (2004) of 10 to 12 participants in cognitive interviews. A purposeful sample of undergraduate and graduate students were selected to represent the breadth of STEM majors, including 4 engineering majors, 2 mathematics majors, 1 psychology major, 2 computer science majors (non-engineering), and 3 STEM education majors (i.e., preservice secondary mathematics and/or science teachers).

Design, instrumentation, and collection of data. For this phase, a concurrent mixed methods design (Creswell, 2002) was employed (see Figure 3.4). The quantitative portion of this mixed methods design was an alignment study assessing content validation/item alignment with content experts using similar methods as employed by Norcini et al. (1993) and Webb (1997). The qualitative portion of this mixed methods design used student cognitive interviews similar to those described by Willis (1999).

The researcher anticipated that subdomains of the STEMSEI should be able to differentiate based on sustainability domain or type of engagement. If the STEMSEI

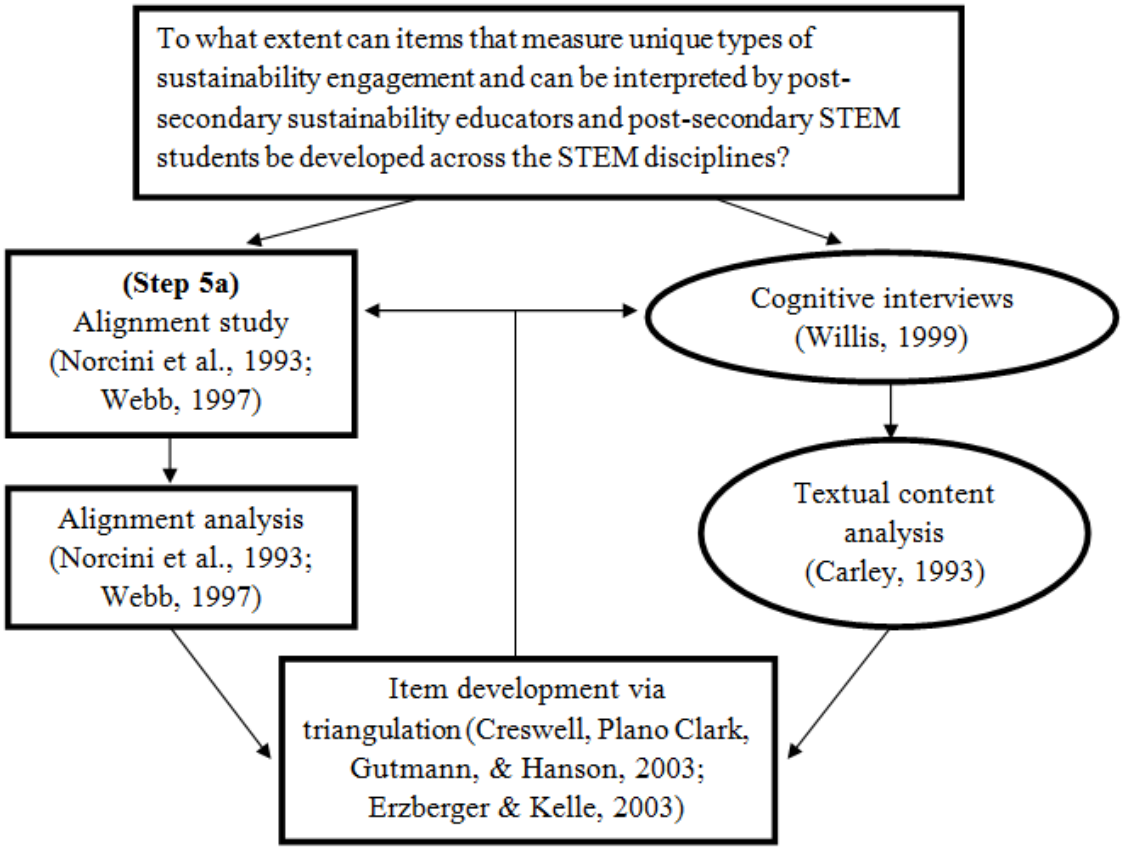

Figure 3.4. Concurrent mixed methods design (Cresswell, 2002) for content validation and item piloting.

differentiated based on sustainability domains, two subdomains were expected: emotional and cognitive engagement. If the STEMSEI differentiated based on engagement type, three subdomains were expected: economic, environmental, and social. In either case, this meant that 84 items should have been developed for the STEMSEI. This is because a subdomain of an instrument should have at least 7 to 10 items (Bohrnstedt, 2010, p. 375) and it is recommended that double the number of items be developed and piloted for an instrument (Benson & Clark, 1982), resulting in 14 items developed for each subdomain. This would have meant 84 items should be developed and piloted (see Table 3.4).

Table 3.4
Items to be Developed Based on Recommendations in the Literature

		Emotional Engagement	Cognitive Engagement	Totals
Sustainability Domains	Environmental	14	14	28
	Economic	14	14	28
	Social	14	14	28
Totals		42	42	84

However, for initial instrument development, the researcher thought this to be too many items to pilot, especially when considering shorter instruments often have higher response rates (Mertens, 2005, p. 196) and that the researcher had no incentive resources to encourage participation. Moreover, since additional rounds were anticipated, the first piloting round could pilot fewer items with additional items piloted subsequent rounds. As such, item development persisted until at least 5 items were determined for each subdomain. Subsequent rounds of piloting would continue until at least 7 to 10 items per subdomain were identified, as recommended by Bohrnstedt (2010, p. 375).

Alignment study methodology. The concurrent mixed methods study design allows multiple rounds of data collection and analysis (Creswell, 2002). The researcher conducted individual, in-person interviews with six content experts in two rounds. In each round, content experts evaluated items in two ways. First, items were assessed for common measurement problems, including readability and interpretability of each item (i.e., respondent validity) (DeMaio & Landreth, 2004). Second, as recommended by Benson and Clark (1982), content experts assigned each item to a particular cell of the

instrument blueprint. The content experts individually assessed the items for alignment to both sustainability domains (economic, environmental, and social) and engagement type (emotional, cognitive, and behavioral). Though the STEMSEI is not designed to measure behavioral engagement, behavioral engagement was assessed by the content experts to ensure no item tapped that particular type of engagement. This consideration was necessary to ensure each item aligned uniquely to one type of engagement. Aligning the items in both sustainability domain and engagement type would provide domain evidence for the STEMSEI items. Since the content experts varied in their specialization in STEM, evaluating the STEMSEI in these two ways would ensure that STEM practitioners could uniformly agree on what a particular item was measuring. This would further ensure the construct validity (see Messick, 1990) of the STEMSEI while making the instrument accessible to sustainability education researchers across the STEM fields. Details of these two evaluation processes follow.

To identify common measurement item problems such as readability and interpretability of each item (i.e., respondent evidence) (see DeMaio & Landreth, 2004), a semi-structured interview protocol was developed (see Appendix C). Content experts were presented a printed copy of the items and asked the questions in the semi-structured interview protocol in Appendix C. Space was provided for the content experts to indicate their responses next to each item. The researcher also had a copy of the printout of items to take notes if the content experts clarified their response in any way. Identifying readability and interpretability problems with items was crucial to ensuring that the target population interprets the items appropriately (DeMaio & Landreth, 2004). This assessed each item's readability/interpretability and alignment to both domain of sustainability (economic, environmental, and social) and engagement type (emotional, cognitive, and/or behavioral). Since each content expert had a different specialization in STEM or a specialization in literacy, completing this process with all the content experts helped ensure each item could be interpreted appropriately across the STEM disciplines (see Appendix A). This was done to ensure the relevance/utility (see Messick, 1990) of the instrument across the diverse population of STEM majors.

Content experts assigned each item to a particular cell of the instrument blueprint (i.e., the alignment study) in two rounds within the interview. For the first round, content

experts assigned each item to a domain or domains of sustainability (economic, environmental, and/or social) that they felt best matched the item's content. Then, the items were reviewed in a second passing and assigned an engagement type or types (emotional, cognitive, and/or behavioral) that best matched the item's content. Content experts were free to assign multiple domains of sustainability or multiple engagement types. That is, an item could be evaluated by content experts to measure both "economic sustainability engagement" and "environmental sustainability engagement." Responses recorded on paper were later transferred to a spreadsheet for analysis. Again, the researcher took any clarifying notes on a separate copy of the items.

If greater than half of the content experts (i.e., a majority) agreed upon both a unique domain of sustainability and a unique type of engagement for an item, that item was considered to appropriately measure a subdomain of the construct of interest (i.e., displaying construct and domain evidence/alignment). Items that did not have a majority of content experts agree on either a unique domain of sustainability or a unique type of engagement were considered for revision or omission. This was similar to the alignment study methodology employed by Webb (1997).

Following feedback from content experts that the method used in the first round was time intensive, a more expedient process was adopted in the second round of the alignment study. The goal was to help save time for the content experts in reviewing the items and to make the process more user-friendly. In an alignment study, Norcini et al. (1993) had content experts use a 5-point Likert scale to judge item's "relevance" to a construct of interest. This scale ranged from "not at all relevant" (= 1) to "very relevant" (= 5). However, this required *a priori* that items be aligned to a unique, specified domain of sustainability (economic, environmental, or social) and a unique, specified type of engagement (emotional or cognitive). After new items were created via editing previous items that were aligned to a subdomain of the construct of interest, the new items were then categorized accordingly in the instrument blueprint matrix (see Table 3.3) by the researcher. When content experts reviewed the items in this round, alignment was judged using a scale similar to that employed by Norcini et al. (1993). Norcini et al.'s (1993) scale was modified to the following for measuring alignment to specified domains of sustainability:

i. Strong Alignment - The item's content ALIGNS STRONGLY with the indicated domain of sustainability (economic, environmental, or social).
ii. Acceptable Alignment - The item's content ALIGNS with the indicated domain of sustainability (economic, environmental, or social).
iii. Insufficient Alignment - The item's content ALIGNS SLIGHTLY with the indicated domain of sustainability (economic, environmental, or social), but other domains are more appropriate.
iv. No Alignment - The item's content DOES NOT ALIGN with the indicated domain of sustainability (economic, environmental, or social).
v. Unable to Judge – The item's content is TOO DIFFICULT TO JUDGE its alignment to the indicated domain of sustainability (economic, environmental, or social).

A similar ranking system was employed for ranking alignment to cognitive or emotional sustainability engagement:

i. Strong Alignment - The item's content ALIGNS STRONGLY with the indicated engagement type (emotional or cognitive).
ii. Acceptable Alignment - The item's content ALIGNS with the indicated engagement type (emotional or cognitive).
iii. Insufficient Alignment - The item's content ALIGNS SLIGHTLY with the indicated engagement type (emotional or cognitive), but other engagement types are more appropriate (emotional, cognitive, or behavioral).
iv. No Alignment - The item's content DOES NOT ALIGN with the indicated engagement type (emotional or cognitive).
v. Unable to Judge – The item's content is TOO DIFFICULT TO JUDGE its alignment to the indicated engagement type (emotional or cognitive).

Using an electronic survey version (Qualtrics) of the items with the response scales above, the content experts separately rated the alignment of items to the domains of sustainability and type of engagement. After all content experts had responded in each round, the data was then downloaded in a spreadsheet and analyzed. An item was determined to fit into the specified cell of the instrument blueprint (see Table 3.3) if a majority of the content experts indicated either a "strong alignment" or an "acceptable alignment" rating for the domain of sustainability and engagement type, both determined a priori.

Cognitive interviews study design. At the same time as the content experts reviewed the items, cognitive interviews were also conducted with 12 post-secondary STEM majors. The cognitive interviews were conducted individually, in person. With respect to validity, cognitive interviews can show empirical evidence for an instrument

while also providing content and construct evidence (Messick, 1990). Both think-alouds and verbal probing are considered cognitive interview techniques (Krosnick & Presser, 2010; Willis, 1999) and are both recommended for cognitive interviews (Krosnick & Presser, 2010; Willis, 1999). Think-alouds are when a respondent reads an item and then verbally communicates their thought processes as they answer the question (Krosnick & Presser, 2010; Willis, 1999). Verbal probing is when the interviewer asks follow-up questions to gather further information or clarification about how the participant responded to a question (Krosnick & Presser, 2010; Willis, 1999). Together, these two processes help researchers identify if an item taps the construct of interest as intended (Krosnick & Presser, 2010; Willis, 1999). Cognitive interviews can also identify construct-irrelevant variance and sometimes its source (Conrad & Blair, 1996). Construct-irrelevant variance degrades measurement accuracy and precision (McMillan & Schumacher, 2010). In short, the cognitive interviews helped ensure that the STEMSEI items were measuring what they were intended to measure (i.e., construct and domain evidence) and that scores produced by the STEMSEI truly measured the construct of interest (i.e., relevance/utility) (see Messick, 1990).

For the 3 rounds of cognitive interviews with post-secondary STEM students, a semi-structured interview protocol (see Appendix D) was developed to assess the ability of the items to accurately measure the construct of interest. Identifying items that do not behave as intended are important to ensure the relevance and utility of the instrument across all STEM majors. Three rounds of cognitive interviews were completed in this study with the same participants each time. During each round of cognitive interviews, participants were shown one item at a time and first asked to "think-aloud" the item (i.e., question 1 of the semi-structured interview protocol in Appendix D). The item was marked for revision using a printout of the items if the participant directly or indirectly mentioned domains of sustainability not targeted by the item (i.e., mentioning environmental issues in a think-aloud of an item targeted towards social sustainability). A similar evaluation for engagement type was also conducted for each item (i.e., implying emotional engagement to respond to an item targeted towards cognitive engagement). Corresponding to the verbal probing portion of the cognitive interview, participants were asked questions 2 and 3 of the semi-structured interview protocol in

Appendix D following the think-aloud. As before, an item was marked for revision if verbal probing responses indicated that the item might measure domains of sustainability or types of engagement not targeted by the item. Corresponding notes for marked items were also taken. Member check (Morse, Barrett, Mayan, Olson, & Spiers, 2002) was used between questions to authenticate cognitive interview data. This was done by repeating the notes taken to the participant to ensure accuracy and modifying said notes when necessary.

Triangulation analysis of marked items. Since there were 2 complete rounds of data gathering from the alignment study and the cognitive interviews, there were 2 rounds of triangulation analysis using data from the alignment study and the cognitive interviews. One final round of data analysis was completed with cognitive interview data, and, hence, was only analyzed using the methods for the qualitative portion of this study (i.e., data was not triangulated). An additional round of expert review was not seen as necessary in the third round since item alignments were particularly strong in the second round (see Chapter IV for details).

Items were added to the item bank if the following criteria were satisfied: (1) no readability/interpretability issues were identified by content experts; (2) a majority of the content experts indicated at least "acceptable" alignment (i.e., either "acceptable" or "strong" alignment) to both a unique sustainability domain and engagement type; and (3) student cognitive interview data indicated no construct-irrelevant variance concerns for that particular item. Items flagged in either the alignment study or the cognitive interviews were further analyzed. Triangulation (see Creswell, Plano Clark, Gutmann, & Hanson, 2003; Erzberger & Kelle, 2003) of the first two rounds of data was conducted after data gathering had closed in each round. Triangulation is a process in which both qualitative and quantitative data are used to develop a more informed conclusion from data (Creswell et al., 2003; Erzberger & Kelle, 2003). For this data, triangulation occurred by comparing similarly marked items from the cognitive interviews and the alignment study. The quantitative data used for triangulation was minimal, and included only the frequency of content experts who assigned an “insufficient alignment”, “no alignment”, or a “unable to judge” response. The qualitative data used for triangulation included the written notes taken by the researcher during the cognitive interviews. If an

item was flagged as misaligning on domain of sustainability in the alignment study and cognitive interview notes indicated a similar issue (i.e., if the quantitative and qualitative data agreed), that item was omitted from the item bank. For items that were flagged on only one portion of the concurrent mixed methods study, item revisions were attempted to resolve the issues identified in either the alignment study or the cognitive interviews.

Phase 3: Instrument Evaluation

At the conclusion of phase 2, an item bank had been produced that tapped the various subdomains of the construct of interest (sustainability engagement). Phase 3 of Benson and Clark's (1982) methodology then evaluated how these items performed within the target population. Benson and Clark's (1982) methodology includes five steps in this phase. However, two steps were modified for this study. Two of Benson and Clark's (1982) steps called for a second piloting and then additional pilot studies if necessary. Instead, this study uses a loop design to account for additional pilot studies beyond the first. Hence, this study uses a three step design for phase 3 (see Figure 3.5).

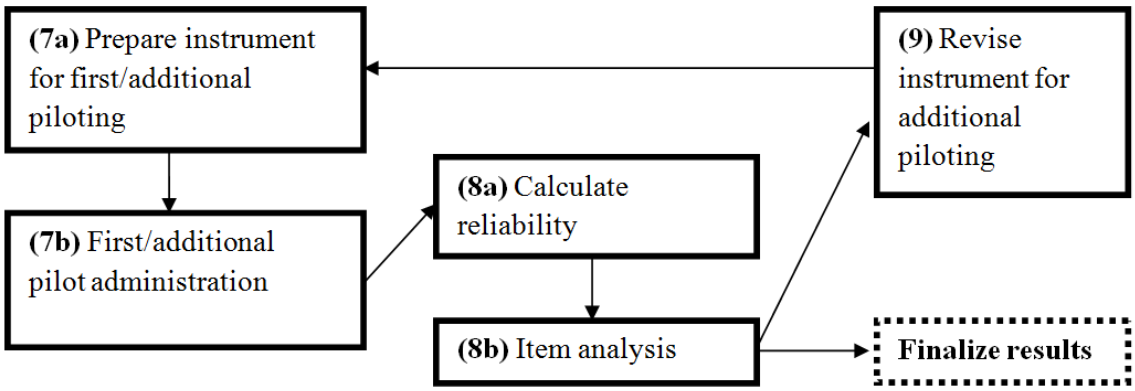

Figure 3.5. Study design of phase 3, instrument evaluation. Individual columns represent different steps in this phase.

Phase 3, step 7a: Preparing for instrument piloting. In step 7a of phase 3, the final items produced in the previous phase needed to be synthesized together in a format conducive for this research. Question order was first considered. First, Krosnick and Presser (2010) note that placing demographic items at the end of an instrument helps reduce respondent fatigue since these items are easy to recall and not sensitive in nature. As such, the demographic items of the STEMSEI were located at the end of the instrument. As another recommendation, Krosnick and Presser (2010) advise presenting items of similar content together as well as placing "easier" items first. With respect to

the domains of sustainability (economic, environmental, and social), there was no clear indication of which items would be "easier". However, the emotional items were seen as "easier" compared to the cognitive items. This was because the emotional items pertained to opinions or preferences while the cognitive questions dealt with meta-cognitive themes (i.e., knowing what your major could do to solve various problems). This meant that the emotional items, or "easier" items, would come first. This was meant to improve respondent motivation and decrease respondent fatigue to the questions (Krosnick & Presser, 2010). Also, since some people prefer certain domains of sustainability over others (Kiewiet & Vos, 2007), there may be a serial order effect (see Krosnick & Presser, 2010) if all of the economic items, environmental items, or social items are grouped together. A serial order effect is when an individual's response on one question impacts responses on latter questions (Krosnick & Presser, 2010). In terms of measurement, this could cause linear dependency in the response vectors. Linear dependency in response vectors can cause measurement errors, such as ill-specified factor structures for an instrument or measurement error (Meyers, Gamst, & Guarino, 2006). In light of this, the emotional items were presented first but were displayed in a random order to all participants. Similarly, all of the cognitive items were presented second but displayed in a random order for each person. Demographic items were displayed at the end in a set order (see Appendices G and H).

After the order of the instrument was decided (the emotional items followed by the cognitive items and finally the demographic items), the item bank of the STEMSEI was uploaded digitally to a Qualtrics server. An electronic medium (i.e., Qualtrics) was selected in this study to disseminate the STEMSEI to (1) allow versatility in collecting data from a sample representative of the target population and (2) to promote ease of use/participation for institutions and participants. Options were set to randomize all of the emotional items and cognitive items in their specified item block. Administration options through Qualtrics were enabled to allow users to complete the survey at a later time and to prohibit multiple responses from individuals. Participants were permitted to skip any question or discontinue at any time. The first and second pilot versions of the STEMSEI can be found in Appendices G and H, respectively.

Phase 3, step 7b: Pilot administration. To complete step 7b (see Figure 3.5), a sample of post-secondary institutions were selected to participate in the pilot studies. Since the STEMSEI was to be calibrated for use across all STEM majors as defined by the National Science Foundation (see Appendix A), undergraduate and graduate students in these majors were targeted across the United States. Originally, a purposeful sample of five United States, doctoral granting post-secondary institutions (PSI) were selected to participate in this study. The PSI's were selected to represent various regions of the United States, which were theorized to reflect the various cultural, political, and sociological factors that might influence measures on the STEMSEI. PSI's were restricted to doctoral granting institutions for two reasons; (1) since the sample was very heterogeneous (post-secondary STEM majors), restricting to this population should help decrease variation in the measurement models and, thus, increase reliability; and (2) PSI's focusing on research (i.e., doctoral granting institutions) were thought to be more likely to use the STEMSEI in sustainability education research, thus the instrument should be calibrated for use in the intended target population.

Student populations at participating institutions varied so as to represent a diverse student population in race, gender, age, and STEM majors. It was also assumed that diversity along past and present religious practices and political views was also present in the student populations at these institutions. Possible faculty liaisons at five post-secondary institutions were emailed a letter inquiring about possible interest in helping administer this study at their home institution (see Appendix E). Three of the five possible faculty liaisons agreed. PSIA was a large research institution (more than 50,000 students) in the South, with PSIB being a smaller research institution (approximately 20,000 students) in the mid-South. PSIC was a smaller research institution (approximately 17,000 students) in the Northeast, while the remaining two non-participating institutions were located in the midwest.

The two latter universities ultimately did not participate due to unforeseen difficulties with administrating study invitations by faculty liaisons at these two institutions. This limits the generalizability of these findings since the sample is dominated by institutions located in the southern geographic region of the United States.

Sample. A total of 764 and 253 post-secondary STEM majors participated in the first and second piloting rounds, respectively. At PSIA, 488 and 193 responses were collected in the first and second piloting rounds, respectively. At PSIB, 214 and 60 responses were gathered in the first and second piloting rounds respectively. The final post-secondary institution, PSIC, only participated in the first piloting round due to institutional research being conducted at that institution during the second piloting round, resulting in 37 responses. For demographic information from the sample, including gender, race, STEM major, and anticipated graduation year, see Table 3.5. Across both rounds, ages ranged from 18 to 62 (M = 20.30, SD = 9.77). The 12 STEM discipline classifications provided in Appendix A are not uniformly represented across universities (i.e., each institution has unique programs and degrees that it offers). Therefore, based on coursework for programs and degrees at each institution, the researcher recoded majors for each participant to the most closely related STEM discipline in Appendix A.

Collection of data. At PSIA, all students majoring in a STEM field (Appendix A) were emailed an invitation to the study (see Appendix F) via an institutionally managed listserv. Members of the information technology team at PSIA sent the invitations via an automated process. Due to technical difficulties, distribution of the invitations was delayed, cutting data collection windows to 3 days out of 14 and 5 days of 7 for the first and second rounds of piloting, respectively. Follow-up emails were not used in either round given the short data gathering window allowed.

At PSIB, data were gathered in two forms: electronically and paper and pencil For the electronically gathered data, Directors of Undergraduate and Graduate Studies of applicable STEM departments were asked to forward the invitation (see Appendix F) to participate in the study to their undergraduate and graduate students. These students were then sent a reminder email 2 weeks after the initial invitation. In order to maximize the number of people who responded, some students majoring in STEM disciplines (see Appendix A) were invited by faculty members in STEM disciplines who were not directors of undergraduate and graduate studies; these students did not receive a reminder email. The electronic data gathering window at PSIB was four weeks during the first round and two weeks for the second round. Finally, paper and pencil data was gathered from students enrolled in a sustainability-related course because the faculty who taught

Table 3.5

Demographic Information for First and Second Piloting Round Participants

Gender	First Round (%)	Second Round (%)
Female	359 (47.0)	118 (46.6)
Male	305 (39.9)	117 (46.2)
Missing	100 (13.1)	18 (7.1)

Race	First Round	Second Round
American Indian/Alaskan Native	3 (0.4)	2 (0.8)
Asian	42 (5.5)	19 (7.5)
Black/African American	11 (1.4)	6 (2.4)
Hispanic	64 (8.4)	26 (10.3)
Native Hawaiin/Pacific Islander	3 (0.4)	0 (0.0)
Middle Eastern	4 (0.5)	3 (1.2)
Mixed Race	10 (1.3)	11 (4.3)
Caucasian	508 (66.5)	168 (66.4)
Missing	119 (15.6)	18 (7.1)

STEM Major	First Round	Second Round
Chemistry	14 (1.8)	5 (2.0)
Computer Science	14 (1.8)	6 (2.4)
Engineering	208 (27.2)	102 (40.3)
Geosciences	58 (7.6)	19 (7.5)
Life Sciences	84 (11.0)	43 (17.0)
Materials Research	0 (0)	0 (0)
Mathematical Sciences	29 (3.8)	7 (2.8)
Physics and Astronomy	12 (1.6)	3 (1.2)
Psychology	32 (4.2)	10 (4.0)
Social Sciences	85 (11.1)	19 (7.5)
STEM Education	46 (6.0)	3 (1.2)
Sustainability Sciences	24 (3.1)	5 (2.0)
Missing	158 (20.7)	31 (12.3)

Anticipated Graduation Year	Undergrad	Grad	Undergrad	Grad
Already Graduated	8 (1.1)	5 (0.7)	0 (0)	2 (0.8)
2014	142 (19.2)	51 (6.9)	27 (10.7)	35 (13.8)
2015	121 (16.4)	41 (5.5)	43 (17.0)	20 (7.9)
2016	121 (16.4)	31 (4.2)	43 (17.0)	21 (8.3)
2017	81 (11.0)	14 (1.9)	28 (11.1)	10 (4.0)
2018	3 (0.4)	10 (1.4)	1 (0.4)	6 (2.4)
2019 or beyond	3 (0.4)	8 (1.1)	1 (0.4)	1 (0.4)
Missing	100 (13.5)		15 (5.9)	

Note. Values in ( ) are the percentages of the sample represented by the corresponding subset of the sample.

the course utilized a pre- post-survey design for their course.

At PSIC, a faculty liaison assisted in targeting specific STEM majors in an effort to maximize faculty support and student participation in the research. Invitations at this institution were sent by either the faculty liaison or faculty members at the participating institution. This institution only participated in the first round due to institutional research being conducted at this institution during round 2. Responses during round 1 were gathered for a total of 2 weeks at this institution, with a reminder sent after 1 week of data collection.

With the exception of a small collection (n = 25) of responses gathered via paper and pencil in round 1 from PSIB, responses were all collected electronically through digital versions (via Qualtrics) of the STEMSEI presented in Appendices H and I. For round 1, the version of the STEMSEI in Appendix H was used while the version in Appendix I was used for round 2. The 25 responses gathered in round 1 at PSIB were collected on paper using the version of the STEMSEI in Appendix H. Those responses were entered by hand into a spreadsheet for analysis.

Since institutional and/or departmental policy varied, the method of invitation to the study could not be made uniform across all participating institutions. Hence, the invitation methods at PSIB and PSIC did not allow an exact response rate to be calculated, although they were estimated to be approximately 4.3% and 7.4%, respectively. The response rate for PSIB was estimated by dividing the number of responses from participants at PSIB by the sum of the estimated number of students in programs that were invited to the study. The estimates of the number of students in programs that were invited to the study were obtained from corresponding departmental websites of PSIB. The response for PSIC was estimated by dividing the number of responses from PSIC by the total number of estimated invitations sent out provided by the faculty liaison at PSIC. The response rate for PSIA was exact due to the automated invitation process via listserv and was calculated to be 3.2%. This response rate was calculated by dividing the total number of responses from participants from PSIA by the exact number of invitations sent out, which was provided by a faculty liaison at PSIA. It was assumed that these emails were all valid, though that could not be confirmed.

Phase 3, step 8a: Calculate reliability. To calculate the reliability for each round of piloting, the researcher used Cronbach's alpha. Cronbach's alpha is an estimate of the internal consistency of the responses on an instrument within a sample. Cronbach's alpha (Cronbach, 1951) is calculated as follows (p. 299):

$$\propto = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) \qquad (1),$$

where n is the sample size, V_i is the variance of item scores after weighting, and V_t is the variance of test scores. In general, Kline (2000) recommends that Cronbach's alpha never be less than .7 (p. 13). For ability tests, Kline (2000) recommends Cronbach's alpha values be around .9 (p. 13). Since the STEMSEI is not an ability test, Cronbach's alpha values greater than .7 were interpreted as sufficient. SPSS 20 (IBM Corp., 2011) was used for all Cronbach's alpha calculations.

Though it could not be calculated at this step, marginal reliability, an analogous statistic to Cronbach’s alpha from item response theory, will be considered later as another estimate of reliability.

Phase 3, step 8b: Item analysis. Benson and Clark (1982) recommend only quantitative evaluation of instruments in their methodology. However, qualitative methods were also considered to improve the methodology in two ways. First, considering additional qualitative data would offer more insights on how to improve the STEMSEI. Second, gathering qualitative data would offer other domain evidence for the instrument. Hence, step 8b of phase 3 utilized a sequential mixed-methods design (Creswell, 2002; Creswell, Plano Clark, Gutmann, & Hanson, 2003) (see Figure 3.6) to analyze the items and establish the STEMSEI's psychometric properties.

In this sequential mixed-methods design, quantitative analyses were performed first, with results indicating that the STEMSEI was differentiating along the sustainability domains (i.e., economic, environmental, and social) (see Chapter IV for details). Quantitative outcomes were then contextualized into a secondary research question (What motivates the distinction between the different domains of sustainability within the population and is it substantiated?). Contextualizing a secondary research question is common practice when using the sequential mixed methods design (Creswell, 2002; Creswell et al., 2003). This transitioned this portion of the sequential mixed-methods

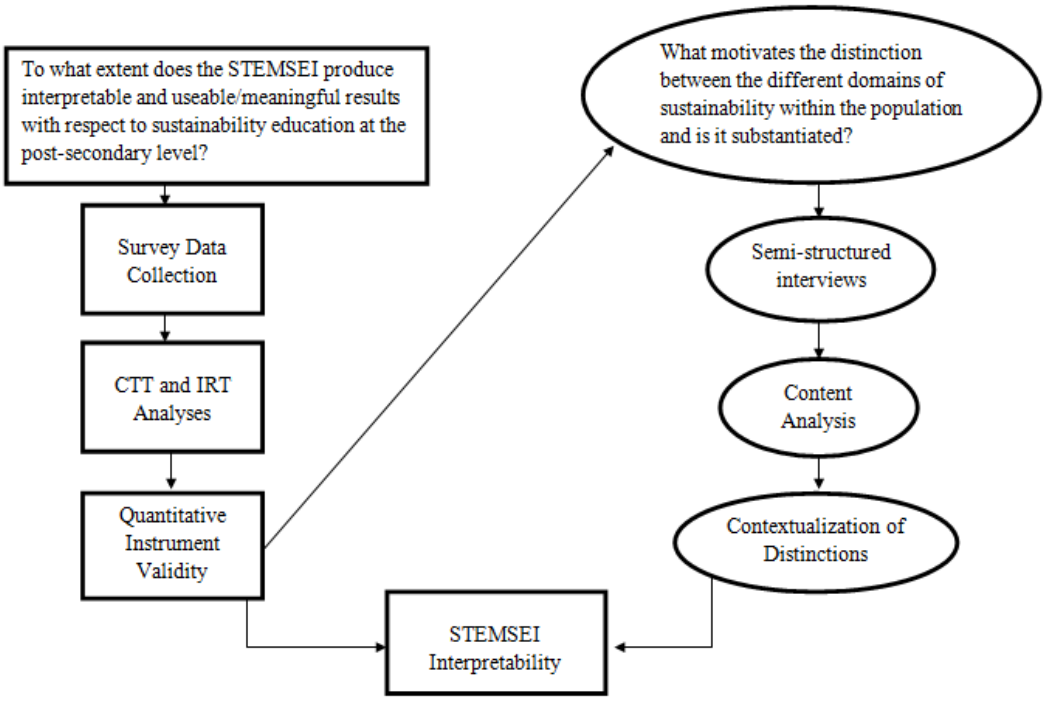

Figure 3.6. Sequential mixed-methods design for STEMSEI piloting. Developed from Creswell (2002) and Creswell, Plano Clark, Gutmann, and Hanson (2003).

study to the qualitative phase. Interviews with post-secondary students were conducted using the semi-structured interview protocol in Appendix F to answer the secondary research question (What motivates the distinction between the different domains of sustainability within the population and is it substantiated?). Triangulation was then used on the quantitative and qualitative data. After items had been edited in light of evidence from both qualitative and quantitative analyses, another round of piloting followed. Triangulation (Creswell et al., 2003; Erzberger & Kelle, 2003) using results from both rounds then guided overall judgment of the STEMSEI's interpretability. See Appendix H and Appendix I for the versions of the STEMSEI used in the first and second piloting rounds, respectively. Two rounds of quantitative piloting were conducted while only one round of qualitative piloting was conducted. Only one round of qualitative piloting was needed since results of the second quantitative piloting round were similar to the first round and did not, therefore, motivate the formation of another secondary research question. In the first round of quantitative piloting, the initial STEMSEI (see Appendix H) was assessed for instrument and item functioning within a sample from the target population. Semi-structured interviews then followed to answer questions arising from the initial quantitative results. Following this, another round of item revisions were

conducted and a second round of piloting followed with the final STEMSEI (see Appendix H).

Sample. The sample for the quantitative piloting (i.e., respondents to the STEMSEI in step 8a of phase 3) (see Figure 3.5) is described in the section for "Phase 3, step 7b: Pilot administration".

A sample of convenience of five post-secondary STEM majors participated in the qualitative portion of this step. The sample included the following; undergraduate engineering majors, both male; a female graduate STEM education major; a female graduate mathematics major; and a male graduate economics major. Participants were invited to this portion of the study directly by the researcher and were initially selected to represent varied ages (from 18 to 28), gender, and classification (undergraduate/graduate student). The initial sample of five post-secondary STEM students was seen as possibly sufficient if qualitative results were convergent. Results of the qualitative pilot were convergent, and the researcher concluded this was an adequate sample given recommendations from McMillan and Schumacher (2010, p. 328).

Quantitative piloting analysis. Since the purpose of this study was to develop an instrument that measured sustainability engagement in the target population, evidence of the generalizability of the instrument and the psychometric strengths and weaknesses of the instrument needed to be determined. Similar to a methodology implemented by Appleton, Christenson, Kim, and Reschly (2006) in their study of student engagement along the subdomains of cognitive and psychological engagement, responses to the first quantitative piloting round were separated into two randomized, approximately equal-sized samples: a development sample and a validation sample. Similar recommendations have been made for item response theory approaches (see De Ayala, 2009).

The development sample is defined as a portion of the sample gathered in step 7b of phase 3 that was used in initial quantitative analyses to assess fit of STEMSEI results to those expected from the theoretical framework developed in phase 1 and phase 2 of this study. The development sample was used to explore the psychometric properties of the STEMSEI The validation sample was the other portion of the sample gathered in step 7b of phase 3 that was used in follow-up quantitative analyses. These follow-up quantitative analyses were confirmatory in nature to see if the data structure that was

found in the development sample could be identified, or confirmed, in the validation sample.

The remaining sections explain the quantitative methodology employed in this study and how the development sample and validation sample were used accordingly.

Factor analyses. Several factor analysis techniques were considered for analyzing the response data to the STEMSEI gathered in step 7b of phase 3. Exploratory and confirmatory factor analysis was considered in the development sample and validation sample, respectively. Overall, there were three factor analysis approaches employed in this study: nonlinear exploratory factor analyses (EFA), nonlinear confirmatory factor analyses (CFA), and nonlinear confirmatory bifactor analyses (CBA). For all factor analyses, fit statistics considered to assess model fit included the chi-square test of model fit, the comparative fit index (CFI), and the root mean square error of approximation (RMSEA). The chi-square test of model fit was interpreted while keeping in mind that this statistic is overpowered for large samples (Meyers, Gamst, & Guarino, 2006). For good model fit, Hu and Bentler (1999) and Yu (2002) recommend CFI ≥ .95. Hu and Bentler (1999) have also recommended RMSEA values close to .06. Similarly, Meyers et al. (2006) have suggested RMSEA values less than .08 to indicate acceptable fit. Multiple fit indices were employed since no one fit index should serve as the sole indicator of model fit and keeping mind that these guidelines are based on continuous variables and may not be applicable to polytomous data. Details of each method and rationale for each follows.

EFA techniques, the first factor analysis technique considered, are appropriate for instrument development or when a factor structure has yet to be identified (Meyers, Gamst, & Guarino, 2006). EFA techniques are also used to discover if responses on an instrument might indicate the presence of subscales or subdomains (Meyers, Gamst, & Guarino, 2006). Testing for the presence of such subscales in responses from the development sample would inform future instrument development. Using response data, EFA's were calculated in M*plus* 6.1 (Muthén & Muthén, 1998-2010) by extracting eigenvalues and corresponding eigenvectors from the interitem polychoric correlation matrix. The eigenvectors represent the "direction" of subscales that are present in an instrument while the corresponding eigenvalues represent how "strong" a corresponding

subscale is in predicting responses on the instrument. An eigenvalue measures how much strength a particular dimension (i.e., the corresponding eigenvector) has in explaining the overall responses given on the instrument. Since M*plus* 6.1 (Muthén & Muthén, 1998-2010) calculates multiple EFA's upon request, the fit statistics detailed above were used to identify EFA solutions with the "best fit".

On the other hand, CFA techniques are appropriate to see if responses to an instrument have a predefined factor structure (Meyers et al., 2006). Predefined factor structures are often established through repeated use of an instrument, such as the factor structure repeatedly identified in responses to the New Ecological Paradigm (NEP) (see Cordano, Welcomer, & Scherer, 2003). CFA techniques can also be used to verify if certain subscales are present in the responses of a particular sample (Meyers et al., 2006). Presence of a specified factor structure in a CFA solution was interpreted through the fit-statistics detailed above.

Confirmatory bifactor analysis (CBA) techniques are similar to CFA techniques in that a predefined factor structure must be given a priori to the analysis. What is different in a CBA, however, is that while the presence of subscales can be identified, the presence of a general factor can still be detected as well (Jennrich & Bentler, 2011a). While some constructs may be multidimensional, they can simultaneously be viewed as unidimensional (Reise, Moore, & Haviland, 2013). Being able to interpret data as simultaneously unidimensional and multidimensional makes the confirmatory bifactor model very useful (Reise et al., 2013). For example, Reise et al. (2013) contend that the bifactor model is useful "for evaluating the plausibility of subscales, determining the extent to which scores reflect a single variable even when the data are multidimensional, and evaluating the feasibility of applying an IRT [item response theory] model" (p. 546). As a specific use of the bifactor model, DeMars (2013) offers that with the bifactor IRT model could allow researchers to control for general scores of mathematics proficiency to study how geometry might be cognitively different from other areas of mathematics (p. 358). In the context of sustainability engagement, a general score on sustainability engagement could control for differences in subdomain areas of sustainability engagement. Furthermore, having a general measure of sustainability engagement might

be helpful for some sustainability educators where the subdomain scores may not be helpful.

There two things the reader should consider with respect to the bifactor models used in this study. First, it would have been favorable to start the bifactor analyses with an exploratory bifactor analysis approach since that approach would more appropriately match the exploratory nature of the analyses in round 1. However, the researcher did not have access to software that could perform exploratory bifactor analyses. On the other hand, the researcher did have access to software that could compute CBA's, which is why they were still considered especially given the utility and versatility that can be derived from bifactor models. The second consideration of the bifactor models used in this study is that they were not orthogonal bifactor models, which is sometimes considered the standard approach (see Reise et al., 2013). In such a model, it specifies a situation where subdomains do not correlate to the corresponding construct of interest (i.e., that the subscale scores do not correlate with the general scale score) (Reise et al., 2013). This has been a common critique of the orthogonal bifactor model since we should expect scores on the general factor to correlate with scores on secondary factors (Reise et al., 2013). However, oblique bifactor models are gaining traction in the literature (see Bandalos & Kopp, 2013; Jennrich & Bentler, 2011b; Reise et al., 2013). Simulation studies with oblique exploratory bifactor models using various oblique rotation methods have been shown to recover original factor structures (Bandalos & Kopp, 2013). Reise and colleagues (2013) explain one such oblique exploratory bifactor model using the Schmid-Leiman orthogonalization (Schmid & Leiman, 1957). The Schmid-Leiman orthogonalization uses an oblique factor rotation (e.g., oblimin) on the interitem polychoric correlation matrix. Though M*plus* 6.1 (Muthén & Muthén, 1998-2010) cannot complete the Schmid-Leiman orthogonalization (Schmid & Leiman, 1957) as an exploratory bifactor analysis approach, this setup can be completed in M*plus* 6.1 (Muthén & Muthén, 1998-2010) as a CBA approach. In such an extraction, one would let the correlations between the general factor and the subfactors be unspecified by the model (i.e., allowed to correlate). For this study, this is how the CBA were conducted in M*plus* 6.1 (Muthén & Muthén, 1998-2010).

As for the rationale for using all three factor analysis methods, EFA's were considered since they can identify the presence of subdomains in an instrument (Meyers, Gamst, & Guarino, 2006), and the content experts had warned in phase 2 that subscales may be present in the cognitive engagement scale: knowing versus solving cognitive engagement. Once a factor structure had been identified by the EFA's, CFA's can then be used to confirm the presence of that factor structure in other samples. This would serve as a validity test to ensure the invariance of the STEMSEI across different samples. Finally, the CBA approach was used to see if a bifactor model could be fit to the data. The bifactor model is of special interest to this research because while it accommodates multidimensional data, it can also produce a univocal measure (Reise, Moore, & Haviland, 2013). In terms of sustainability engagement, this would mean that an overall measure of sustainability engagement could be attained, which may be of particular use to sustainability educators.

In round 1, all three techniques were utilized. First, an EFA was conducted using responses from the development sample of round 1. Using the factor structure identified in the EFA's, a CFA approach was then used to confirm the presence of the indicated factor structure in the validation sample's responses to the STEMSEI. This served as a validity check to ensure the factor structure of the instrument was maintained across different samples (i.e., that the factor structure was independent of the sample). Following the CFA, a CBA was conducted also using the same factor structure indicated by the EFA, though recall that the CBA also estimates a general, univocal measure (Reise, Moore, & Haviland, 2013). In round 2, a confirmatory approach was taken that reflected results from round 1. That is, either a CFA or CBA will be performed in round 2 depending on results from round 1. The purpose of this is to see if the structure of the STEMSEI is invariant across the rounds despite additional items.

All three factor analysis techniques were performed in M*plus* 6.1 (Muthén & Muthén, 1998-2010) using the logit link function or polychoric correlations (Olsson, 1979). In factor analysis extractions, polychoric correlations are preferred over Pearson's correlations (Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2010) because unlike Pearson's correlations, polychoric correlations do not attenuate the estimation of ordinal data that approximates a continuum (Holgado–Tello et al., 2010).

Due to this lack of attenuation, polychoric correlations function better at recovering original factor structures in EFA (Holgado–Tello et al., 2010). M*plus* 6.1 (Muthén & Muthén, 1998-2010) was also used to conduct an oblimin rotation for each EFA with the mean and variance adjusted weighted least squares estimation (WLSMV) method. Since orthogonal rotations remove any correlation between factors (Meyers, Gamst, & Guarino, 2006), an oblimin rotation was preferred over all orthogonal extraction techniques since correlation between factors were expected. For all factor analyses, if an item cross-loaded onto multiple factors after an oblimin rotation (i.e., the factor pattern coefficient is large for multiple factors), the item was omitted since a simple structure or solution was desired (i.e., items load uniquely onto one and only one factor). Using criteria recommended by Matsunaga (2010), an item is considered to be cross-loading if the factor pattern coefficients exceed .3 for at least 2 factors. For example, if an item's pattern coefficients are .2, .2, and .6 for a 3-factor solution, then this item would not be considered to be cross-loading. On the other hand, factor pattern coefficients of .2, .4, and .8 would be considered cross-loading.

Finally, three factor structures were seen as possible. First, a unidimensional factor structure (i.e., viewing sustainability engagement as one cohesive factor) (see Meadows, 2008; World Commission on Environment and Development, 1987) was possible. Content expert feedback from phase 1 also supported this idea (see Chapter IV). Second, a multidimensional factor structure similar to that of the quantifiable definition of sustainability was also possible. In this case, it was expected that items of similar sustainability domain load onto common factors (i.e., economic, environmental, and social items loading separately onto corresponding factors). Third, another multidimensional factor structure similar to that of the sustainability engagement framework was also plausible. In such a case, items of similar sustainability engagement would load onto separate corresponding factors (i.e., an emotional engagement factor and a cognitive engagement factor). One point to consider is that interitem correlations between the six types of items reflected in the instrument blueprint (see Table 3.3) were expected. For example, emotional economic engagement items were expected to correlate with cognitive social engagement items. In this case, factor analyses may have to be completed on a reduced set of items to reduce noise in factor extractions.

Item response theory analyses. For this section, definitions for terms are provided first for reader ease. In both item response theory (IRT) models and Rasch models, person locations are estimates of where a person falls on the latent trait continuum, which is usually considered to be the interval [-3, 3]. Item parameters, which includes item locations and item slopes, are values which are estimated for each item so as to improve model fit. Item slopes are represent how well an item can differentiate between respondents located on different ends of the latent trait continuum (De Ayala, 2009). Item locations represent where an item is located on the latent trait continuum (De Ayala, 2009). Item location values are interpreted differently based on the model that is considered. These will be explained in more detail later.

When looking at possible models to use in this study, there are proponents for both item response theory (IRT) models and Rasch models. However, Rasch models were not considered for this study because of a philosophical difference in how Rasch models and IRT models estimate person and item locations. Rasch models are seen as "the standard by which one can create an instrument for measuring a variable" (p. 19, de Ayala, 2009). Rasch models define a unit of measurement and estimate person and item values in a way that optimizes data-fit (i.e., makes the data fit the model). IRT models do not fit person and item estimates in this way. IRT models try to make the model fit the data. Since the researcher ascribes to the philosophy behind IRT models (i.e., not making the data fit the model), only IRT models were considered in this study.

After the factor structure of the responses to the STEMSEI was identified for the development sample, IRT models were estimated. Initially, IRT models were calibrated for the development sample. Using the factor structure indicated from the factor analyses, comparable IRT models were considered for the development sample. Matching the factor structure of the IRT model to the factor structure of the latent trait is important because of the dimensionality assumption of all IRT models. If a construct is truly unidimensional, as revealed or confirmed by factor analyses, then a unidimensional IRT model should be employed since the underlying latent trait continuum is unidimensional (de Ayala, 2009). On the other hand, if a construct is multidimensional, as revealed or confirmed by factor analyses, then a correspondingly structured multidimensional IRT model should be employed since the underlying latent trait

continuum is multidimensional (de Ayala, 2009). If results from the factor analysis indicated multidimensionality in the data, then only multidimensional models were considered. Also, since a polytomous response scale was selected for all items in step 3b of phase 1 (see Figure 3.1), only polytomous item response models were considered. Concerning multidimensional, polytomous IRT models, two such models were considered: the Multidimensional Generalized Partial Credit (MGPC; Reckase, 2009, p. 102) Model and the Multidimensional Graded Response (MGR; Reckase, 2009, p. 107) Model. One other multidimensional, polytomous IRT model that was considered was the bifactor IRT model (see DeMars, 2013). Again, this multidimensional, polytomous IRT model was also considered due to the model's ability to produce univocal scores for multidimensional constructs (Reise et al., 2013). Such univocal measures would be beneficial to users of the STEMSEI and sustainability educators

First, a description of the MGPC model is presented from Reckase (2009). The MGPC model is a multidimensional extension of the generalized partial credit model (Muraki, 1992). The MGPC model generates an estimate of a person's location in the latent trait space (i.e., more than one latent trait continuum) and is explicitly described by the following mathematical equation (Reckase, 2009, p. 103):

$$P\left(u_{ij} = k \middle| \theta_j\right) = \frac{exp\left(ka_i\theta_j' - \sum_{u=0}^{k} \beta_{iu}\right)}{\sum_{v=0}^{K_i} exp\left(va_i\theta_j' - \sum_{u=0}^{v} \beta_{iu}\right)} \quad (2)$$

A person's location in the *n*-dimensional latent trait space is a vector, θ, that is composed of *n* components. The value in component-*j* of the vector θ, or θ_j , represents the individual's location on the corresponding latent trait continuum with dimension *j*. Due to the multidimensionality that is assumed in the MGPC model, an item slope vector, a_i, is usually estimated for each item *i* (Reckase, 2009). In this case, however, the item slope vectors are essentially reduced to scalars since each item will load onto only one dimension of the IRT model (i.e., for each item, all but one component of the item slope vector will be 0). This is because after the oblimin rotation described in the factor analysis section, items that cross-load (i.e., contribute to multiple factors) are omitted from STEMSEI. Hence, we can assume that each item loads uniquely onto one dimension. For each adjacent pair of response options, the MGPC model also estimates

threshold parameters, β_{iu}, with $\beta_{i0} = 0$ by definition. These values are akin to measurements of item difficulty and represent the points at which a respondent is equally likely to be assigned a score from the lower category or the upper category (i.e., it is where two adjacent category probability curves intersect). Also, assuming an instrument has $K_i + 1$ score categories overall, K_i represents the maximum score for item i, where $0 \leq k \leq K_i$.

Now, consider the multidimensional graded response (MGR) model. Reckase (2009) describes the MGR model in terms of the normal ogive form, which is not the form we will use with IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011). Instead, de Ayala's (1994) version of the MGR model using logistic curves is presented here, which is compatible with IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011). The mathematical equation for the MGR model is given as (De Ayala, 1994, p. 156):

$$P_{x_i}(\theta) = \frac{exp[D \sum a_{ih}(\theta_h - d_{xi})]}{1 + exp[D \sum a_{ih}(\theta_h - d_{xi})]} \qquad (3)$$

As before, a person's location on the n-dimensional latent trait space is given by the vector θ, where component-h of the vector θ , or θ_h, represents the person's location on the latent trait continuum corresponding to dimension h. In the MGR model, a_{ih} measures the item slope parameter of item i along dimension h; that is, in general, unique slope parameters are estimated for each item across each dimension. As before, though, the item slope vectors are essentially reduced to scalars since each item will load onto only one dimension of the IRT model (i.e., for each item, all but one a_{ih} will be 0). Next, m_i represents the number of ordered category boundaries (i.e., there are $m_i + 1$ ordered categories), and x_i varies between any of the ordered categories (i.e., $0 \leq x_i \leq m_i$). For each item i, unique thresholds are estimated, d_{xi}, for each item (de Ayala, 1994).
A scaling factor of $D = 1.702$ is also used in this model. De Ayala (1994) further explains:

> $P_{x_i}(\theta)$ is the probability of a randomly selected examinee with latent traits θ responding in category x or higher for item i. The probability of responding in the lowest category (i.e., P_0) or higher is defined as 1.0, and the probability of responding in the highest category (i.e., P_{m_i+1}) is 0.0. For example, for an item with four response categories (i.e., 0, 1, 2, and 3), $P_2(\theta)$ is the probability of responding in categories 2 or 3 rather than in categories 0 or 1. (p. 156)

Finally, an explanation of the bifactor model is given. In equation form, the bifactor model is given by (DeMars, 2013, p. 377):

$$P(x = x_j) = \int \prod_{s=1}^{S} \left[\int L_{js}(\theta_j, \theta_s) g(\theta_s) d(\theta_s) \right] g(\theta_g) d(\theta_g) \qquad (4)$$

To maximize probabilistic modeling of the general factor, the bifactor model is marginalized over θ_g. DeMars (2013) explains that the bifactor model is similarly estimated as the other models explained above. However, instead of allowing an item to load only on one dimension (as in the MGR and MGPC models above) or onto all dimensions (as is the case with generalized multidimensional models), each item is allowed only to load onto the general factor and a unique secondary factor (DeMars, 2013). Suppose a model has S secondary factors. So, for any person location in the latent trait space, there would be a person location in the general latent trait continuum, θ_g, and a person location in each of the secondary latent trait continuums, θ_1 , $\theta_2, \ldots, \theta_S$. First, for $1 \leq s \leq S$, each secondary factor s is individually marginalized as a function of θ_g and θ_s , or L_{js}, to maximize probabilistic modeling over each of the secondary factors (DeMars, 2013). Then, the likelihood function of the general factor, or L_j, is marginalized over each of the secondary factors (DeMars, 2013).

Sometimes when item parameters are estimated, item slope parameters can be outside the tolerable bounds of .5 to 3 (Baker, 2001). Recall that the item slope parameters are a measurement of an item's ability to discriminate between different respondents' locations in the latent trait space. While it may seem better to have infinitely large values for the item slopes to provide more discrimination power, this is not true. The reasoning behind this is that one can only reasonably expect an item to be able to discriminate within the sample to a certain degree. In cases where item slope parameters are outside tolerable bounds, prior distributions for item parameters are implemented (see Matteucci, Mignani, & Veldkamp, 2012) to get stable estimates. For this study, a prior normal distribution ($\mu = 1.7$, $\sigma = 0.5$) was implemented in such cases, similar to methods used by Matteucci, Mignani, and Veldkamp (2012).

Since scales along the sustainability domains would likely be most helpful to sustainability educators, IRT models were calculated in this fashion. That is, an economic scale, environmental scale, and social scale were created using the

corresponding items for each scale. Moreover, the dimensional structure of the multivariate IRT models was made similar to the corresponding factor structures indicated in the factor analyses of the previous section. For example, the MGR model for the economic items used 3 dimensions each corresponding to the emotional, cognitive (knowing), and cognitive (solving) items. Similarly, a bifactor IRT model for the environmental items was used which consisted of 4 dimensions; one for the general dimension, and then one each for the emotional, cognitive (knowing), and cognitive (solving) items. All IRT analyses and models were calculated in IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011).

For each model, items were then assessed for local dependency using the standardized local dependency (LD) χ^2 statistic (Chen & Thissen, 1997). Local dependency is first tested to ensure that the local independence assumption of all IRT models is met. Essentially, the local independence assumption of IRT models is that "how a person responds to a question is determined solely by his or her location on the latent continuum and not by how he or she responds to any other question" (de Ayala, 2009, p. 20). This assumption is sometimes seen as irrevocably tied to the dimensionality assumption of all IRT models as well (de Ayala, 2009). If the standardized LD χ^2 statistic was 10 or greater for an item pair, one of the items were considered for removal, as suggested by Cai, du Toit, and Thissen (2011, p. 77). To determine if an item was to be removed in this case, the model was recalibrated without the item in question and the item parameters were compared between the two calibrations; the item was not removed if the item parameters remained mostly unchanged between the two calibrations. After removal of an item, the corresponding model was reestimated and the standardized LD χ^2 statistics were reevaluated. This process repeated until there were no concerns of local dependency as evidenced by the LD χ^2 statistic.

Once no item pairs exhibited LD, item-level fit was assessed in the development sample IRT models using the S-χ^2 item level diagnostic statistic (Orlando & Thissen, 2000, 2003). The S-χ^2 item level diagnostic statistic (Orlando & Thissen, 2000, 2003) estimates how well each item fits to model expectation (Orlando & Thissen, 2000). Following a Benjamini-Hochberg correction (see Benjamini & Hochberg, 1995) to control for inflated Type 1 errors, items with a statistically significant ($\alpha = .05$) S-χ^2 item

level diagnostic statistic were further examined at the item level. For such items, MODFIT 3.0 (Stark, 2008) was used to assess item fit graphically. Using the response data used to generate item parameters for an IRT model as well as the item parameters themselves, MODFIT 3.0 (Stark, 2008) creates item-fit plots that displays both the option response curve and the empirical response curve for each response option for each item. The option response curve is the probabilistic logistic curve used to predict option response based on a person's location on the latent trait continuum (de Ayala, 2009). The empirical response curve is the empirical probability of a person responding a certain way to an item given their location on the latent trait continuum. The option response curve and the empirical response curve should overlap or be very similar to one another (i.e., overlap to a certain degree). This indicates that the model prediction is close to what actually occurs in the sample. If there is a large degree of misfit between these curves, it indicates a lack of item-level fit. Misfit was judged by ensuring that the empirical probability curve was within the 95% confidence interval of the option response curve. Only items that returned a statistically significant *p*-value from the S-χ^2 item level diagnostic statistic and exhibited poor item-fit in the fit-plots generated by MODFIT 3.0 (Stark, 2008) as described above were removed. When an item was removed in this fashion, the corresponding IRT model was reestimated. This process was repeated until item-level fit was found for all items using the criteria outlined above.

Next, the IRT models for the development sample were compared to see which model offered the best fit. Using Akaike's Information Criteria (AIC) statistic (Akaike, 1987), non-nested models, such as the models used in this study, can be compared using the AIC difference Test (Burnham & Anderson, 2002). The AIC Difference Test (Burnham & Anderson, 2002) calculates the absolute value of the difference between two models' AIC statistics. If the AIC Difference Statistic for two models is between 0 and 2, there is substantial support for both models (Burnham & Anderson, 2002, p. 70). On the other hand, if the difference statistic is between 4 and 7, the model with the smaller AIC statistic is considered slightly favorable to the other model (Burnham & Anderson, 2002, p. 70). Finally, if the difference statistic is greater than 10, the model with the higher AIC is not considered in further analyses (Burnham & Anderson, 2002, p. 71). These

guidelines were used to identify which IRT models should be further refined to improve model-fit.

Final IRT models were also assessed for marginal reliability and the standard error of estimate (SEE). Marginal reliability is a measure of the precision of the person location estimate. It is sometimes considered analogous to Cronbach's alpha (de Ayala, 2009) and is calculated as (de Ayala, 2009, p. 205):

$$\rho = \frac{\sigma_\theta^2 - \sigma_{em}^2}{\sigma_\theta^2} \qquad (5)$$

where σ_θ^2 is the variance of the person location estimates and σ_{em}^2 is the variance of the error of measurement. In this study, the person location estimates were generated using a normal distribution (μ = 0, σ = 1), so $\sigma_\theta^2 = 1$ by assumption for the purposes of this study. Moreover, σ_{em}^2 is the SEE. Marginal reliability values closer to 1 indicate higher accuracy of the person location estimate while values closer to 0 indicate little to no accuracy of the person location estimate (de Ayala, 2009). Similarly, the SEE is a measure of the estimated error of the person location estimates (de Ayala, 2009) and can vary across the latent trait continuum, meaning that the accuracy of scores produced by the instrument can vary across the continuum. This is because the total information provided by an instrument varies across the continuum as well. Regardless, the SEE should be close to 0 across the continuum (de Ayala, 2009).

Validation assessments and hypotheses. Since the STEMSEI is a new instrument, multiple validity assessments were considered.

First, differential item functioning (DIF) analyses were considered for each scale. A DIF analysis identifies items in an IRT model that are functioning differently across various subsamples even after controlling for person location on the latent trait continuum (de Ayala, 2009). Items exhibiting DIF indicates that the indicated items are not invariant across subsamples. Ultimately, an instrument should be invariant across subsamples (i.e., no item favors or advantages one group over another).

When an item exhibits DIF, there is a statistically significant difference in item parameters estimated for one group compared to those of another group even after controlling for person location on the latent trait continuum. This is of particular concern for instruments because such differences indicate one group is disadvantaged on a

particular item when compared to another group (de Ayala, 2009). Moreover, such differences can impact estimated person locations by the IRT models (de Ayala, 2009), which could then artificially create group differences (i.e., statistically significant differences between group means of person location on the latent trait continuum). DIF is also an indicator of possible bias in an item (de Ayala, 2009), but such bias is usually confirmed with a panel of content experts (Kararni, 2012).

Unfortunately, IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011) cannot compute DIF analyses for multidimensional IRT models, which were used to score responses to the STEMSEI. As an attempt to still assess the items for DIF, a unidimensional graded response model (see Samejima, 1969) for each sustainability domain was estimated in IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011) and subsequently assessed for DIF. Despite the lack of fit between the unidimensional DIF analysis techniques available to the researcher and the multidimensional IRT models employed in the STEMSEI, the DIF analyses were pursued because of the effects of item bias on scores. One such difference includes introducing artificial group mean differences.

Two rounds of DIF analyses were implemented in IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011). IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011) uses Wald tests (Lord, 1980), a type of chi-square test, for all DIF analyses. Three chi-square test statistics are produced for any DIF analysis: an omnibus statistic, a uniform DIF statistic, and a non-uniform DIF statistic. The uniform DIF statistic and non-uniform DIF statistic are only interpreted if the omnibus statistic is statistically significant at the $\alpha = .05$ level. To control for inflated Type 1 error rates and to make the DIF analyses more conservative given the mismatch of the technique to the IRT models used, a Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) was applied to all DIF analyses. The initial DIF analysis assessed all items on each scale for DIF as described above.

If the initial DIF analyses indicated possible presence of DIF, a secondary DIF analysis was performed. In the secondary DIF analyses, items not displaying DIF in the initial DIF analysis were anchored (i.e., were set as anchor items) and then the DIF analyses were recalculated. An anchor item is an item that does not appear to have DIF, as determined by the initial DIF analyses. The secondary DIF analyses were then

interpreted the same as the initial DIF analyses using the three chi-square statistics as described above. Items displaying statistically significant DIF in the secondary DIF analyses were deleted. Normally, such items would be reviewed for bias by a panel of experts (Kararni, 2012), but such a panel was not available to the researcher at the time.

A statistically significant result on the uniform DIF statistic means estimated item thresholds/locations between the two groups are significantly different statistically (de Ayala, 2009). For example, consider gender. Uniform DIF with respect to males (focal group) would mean a male indicating the same response as a female (reference group) would score either higher or lower on the latent trait continuum. Equivalent responses do not mean equivalent locations on the latent trait continuum for these groups. For a particular location on the latent trait continuum, the option response curves can be used to determine which group has the "advantage" on this item. If the option response curve for the focal group is below the option response curve for the reference group, the focal group finds it harder to endorse that particular response on the item. In such a case, the estimated person location for the focal group should be greater than that estimated for those in the reference group because of the "disadvantage" the focal group has with respect to the item. The opposite is true if the option response curves for the focal group are above the option response curves for the reference group. In uniform DIF, option response curves between two groups are translations of one another.

On the other hand, an item showing statistically significant non-uniform DIF means the estimated item slope parameters for the focal group is significantly different statistically than that of the reference group (de Ayala, 2009). A non-uniform DIF item provides more information of the person location on one part of the latent trait continuum while providing less information on other parts (de Ayala, 2009).

There are two final considerations to the DIF analyses. First, any statistically significant results in the DIF analyses were interpreted with caution. This is because of the mismatch between the methods of DIF analysis available to the researcher and the IRT models used in the STEMSEI. Second, demographic variables needed to be chosen to delineate possible subgroups where DIF might be present in the STEMSEI. Given that the STEMSEI is a new instrument, no prior DIF results existed for this instrument. However, using prior research in fields closely related to the construct of interest (i.e.,

sustainability engagement), some variables were identified to consider for the DIF analyses. The researcher found no prior literature related to the scope of this study that established statistically significant differences on affective economic or affective social scales based on demographic variables (i.e., gender, race, age, etc.). However, DIF was still assessed over race, gender, religion, and classification for the economic and social items. On the other hand, researchers have identified some demographic variables that impact measures on affective environmental scales (refer to the "Expected outcomes along environmental engagement" in Chapter III for an in-depth discussion of these).

Some DIF analyses could not be performed with the original STEMSEI development sample. This was because of sample size limitations. For example, DIF analyses in simulation studies using the Graded Response Model (Samejima, 1969) supplied meaningful results provided sample size per group was at least 250 (Langer, 2008). In similar sample sizes per group, type I error rates for detecting DIF at the .05 level ranged from .01 to .03 (Langer, 2008). Considering this, some DIF analyses along the gender, race, and religion variables would be underpowered. Considering a DIF analysis along race, since there were only 11 African American respondents in the first round (6 of which were in the development sample by random assignment), a DIF analysis with African American students as the focal group would not be reasonable. Hence, as discussed in the "Expected outcomes along environmental engagement" section in Chapter III, race was recoded as either "white" (= 0) or "student of color" (= 1). For the purposes of this study, students of color included Native Americans, Alaska Natives, Asians, Black/African Americans, Hispanics, Native Hawaiian/Pacific Islanders, and Middle Easterners. Similar problems were encountered with the religion variable since a variety of religions were indicated by the respondents. For similar reasons discussed in the "Expected outcomes along environmental engagement" section, past and present religion was recoded as either Christian (= 0) or non-Christian (= 1).

Second, another validity test for IRT models is invariance of estimated item parameters. Morizot, Ainsworth, and Reise (2009) explain that the item parameters for an IRT model should not depend on the sample from which they are estimated from. Hence, for any instrument, item parameters generated from two different samples should be very similar (de Ayala, 2009; Morizot, Ainsworth, & Reise, 2009). To test for this

with the STEMSEI, item parameters for all final IRT models were estimated using responses from the validation sample. These item parameters were then compared to those estimated for the development sample. A large, positive correlation was expected, which would indicate model invariance at a linear level (de Ayala, 2009). This would mean the STEMSEI would provide similar item parameter estimates even if different samples were used, showing invariance and providing initial evidence of generalizability for the instrument to those being compared (Morizot, Ainsworth, & Reise, 2009).

A third validity test was the invariance of person location scores on the latent trait continuum when using different item parameters. If one estimates two different sets of item parameters using responses from two different samples (i.e., the development sample and the validation sample), one should expect that using either set of item parameters, person location estimates for either sample should be invariant (Morizot, Ainsworth, & Reise, 2009). This is because a person's location on the latent trait continuum should be independent of the items (Morizot, Ainsworth, & Reise, 2009). For all final models, a second person location estimate was generated for each respondent in the development sample using the item parameters estimated from the validation sample responses. Then, the two person location estimates were correlated. A large, positive correlation was expected, which would indicate model invariance (Morizot, Ainsworth, & Reise, 2009) at the linear level or that they are consistent.

A fourth validity test for invariance, the root mean squared difference (RMSD) test statistic (de Ayala, 2009), was also considered. For a given interval on the latent trait continuum (e.g., [-3, 3]), the RMSD test statistic estimates the average difference between corresponding option response functions of two different models (de Ayala, 2009, p. 113):

$$RMSD_j = \sqrt{\frac{\sum(p_{j_1} - p_{j_2})^2}{n}} \qquad (6)$$

where p_{j_1} and p_{j_2} are corresponding option response function values along the continuum and *n* is the number of subintervals used in the estimation. The RMSD should be small (close to zero), indicating little difference in the option response functions, and thus the overall models (de Ayala, 2009).

A fifth test of invariance was considered. Since IRT models have stronger model-fit the sample in which they are calibrated, the validation sample was scored using item parameters estimated from the development sample. That is, using the item parameters estimated from the development sample, empirical model-fit of the responses of the validation sample was assessed using MODFIT 3.0 (Stark, 2008). The graphical plots generated by MODFIT 3.0 (Stark, 2008) were assessed for model-fit just as they were for the item-level-fit analyses described in the previous section. If item-level fit can be found across samples like this (i.e., mixing item parameters estimated from one sample with the responses on another), this would provide evidence of the invariance of the STEMSEI across different samples, thus providing evidence of generalizability.

Finally, along certain variables some statistically significant differences in the STEMSEI environmental scores were expected. These included gender, race, and religion. Statistically significant differences due to STEM major were also expected due to distinctions in traditional STEM disciplinary focuses (e.g., geosciences majors should be more engaged in environmental sustainability than other STEM majors, while the social sciences majors should be more engaged in social sustainability than other STEM majors). On the other hand, differences between scores from undergraduate and graduate students were not expected.

To assess for differences between groups, MANOVA's were first considered. The MANOVA's were performed across each scale (i.e., economic, environmental, and social) since each scale is a separate IRT model (i.e., technically three distinct instruments). Meyers et al. (2006) recommend that the number of cases per cell of a MANOVA test exceed the number of dependent variables being assessed (p. 375), which was ensured for all MANOVA tests. However, this meant that a non-factorial design had to be employed for the MANOVA analyses, because in a full factorial model several cells would have been empty. Further rationale for this is rooted in the fact that a MANOVA determines the combined weighted linear composites to maximally distinguish among the variates (Meyers et al., 2006, p. 366). A factorial design attributes some of the variation to interaction effects (Meyers et al., 2006, p. 441). This means that non-factorial design would maximize the variates between one set of groups (i.e., not among two or more groups where the variates are composites of two or more variables). The argument here

is that maximizing differences for each variable will help with the development of the STEMSEI by allowing each demographic variable to be analyzed individually and allowing the maximum amount of variance to be attributed to that demographic variable. Due to the non-factorial design, there were five separate MANOVA analyses (i.e., one for race, gender, religion, classification, and STEM major) for the economic, environmental, and social scales, separately.

Post-hoc ANOVA tests were conducted when statistically significant MANOVA's were indicated except for STEM major. The rationale for omitting STEM majors from post-hoc tests is because it is just the fact that the STEMSEI can differentiate along STEM majors that is important to establish (as validity evidence). Post-hoc ANOVA tests for race, gender, religion, and classification (if applicable) were assessed at the $\alpha = .01$ level as a way to control for Type 1 errors while also applying a Benjamini-Hochberg correction. For statistically significant post-hoc ANOVA tests, differences between groups were then assessed using independent t-tests also at the $\alpha = .01$ level to control for Type 1 errors. Where applicable, effect size will be measured with Cohen's d and the partial eta squared.

Synopsis

This chapter presented the methodology for this study, outlining the three phases of this study: instrument planning, instrument construction, and instrument evaluation. Within each phase, a study was conducted to adhere to the recommendations of Benson and Clark (1982) for the various steps of instrument development. In phase 1, a qualitative study sought to determine the extent that a convergent theoretical framework for sustainability engagement in post-secondary STEM students could be achieved 6 content experts in the field. In phase 2, a concurrent mixed-methods study was conducted to determine the extent to which items that measure unique types of sustainability engagement can be developed so that they are interpretable by post-secondary sustainability educators and post-secondary STEM students. Finally, in phase 3, a sequential mixed-methods study was conducted to determine the extent to which the STEMSEI provided interpretable and useful/meaningful results with respect to post-secondary sustainability education.

Copyright © David L. Little II 2014

Chapter IV

Results

This chapter presents the results across all three phases of this study. Results are presented in the order in which each part was conducted. As such, presentation of results will follow the 9-step methodology adapted from Benson and Clark (1982) as outlined in the previous chapter.

Some steps in the methodology used in this study do not have clear results that can be presented for the reader. For example, step 4b was to train item writers (see Figure 3.3). The results of this step are seen throughout the rest of the study and can be indirectly observed in the initial and final STEMSEI (see Appendices G and H). Hence, results for such sections will not be directly provided to the reader.

Also, the methodology adopted from Benson and Clark (1982) is sequential in nature with interpretation of results at one step necessary before continuing to another step. Consider step 8b of phase 3 (item analysis). This step required that interpretation of the EFA results occurred before transitioning to IRT analyses since appropriate IRT models cannot be chosen before knowing the factor structure of the instrument. Hence, interpretation of results is presented where needed to assist the reader in understanding the researcher's progress through the methodology.

Phase 1: Instrument Planning Results

The purpose of this phase of the study to plan the STEMSEI's design and format. In that, the theoretical framework of the STEMSE was edited, evaluated, and validated with content experts in sustainability using the Delphi method (Okoli & Pawlowski, 2003). For reader ease, the study diagram for this phase is presented again here (see Figure 4.1).

Phase 1, step 1 results: STEMSEI purpose, target population, and domains. The purpose of the STEMSEI was to measure sustainability engagement (such as emotions and opinions towards sustainability as well as knowledge of and development of solutions for sustainability) in post-secondary STEM majors. Hence, the population of the STEMSEI is post-secondary STEM majors (see Appendix A). The domains and subdomains of sustainability engagement are measured by the STEMSEI include the following:

- Economic sustainability engagement (a general scale)
    - Emotional economic sustainability engagement
    - Cognitive (knowing) economic sustainability engagement
    - Cognitive (solving) economic sustainability engagement
- Environmental sustainability engagement (a general scale)
    - Emotional economic sustainability engagement
    - Cognitive (knowing) economic sustainability engagement
    - Cognitive (solving) economic sustainability engagement
- Social sustainability engagement (a general scale)
    - Emotional economic sustainability engagement
    - Cognitive (knowing) economic sustainability engagement
    - Cognitive (solving) economic sustainability engagement

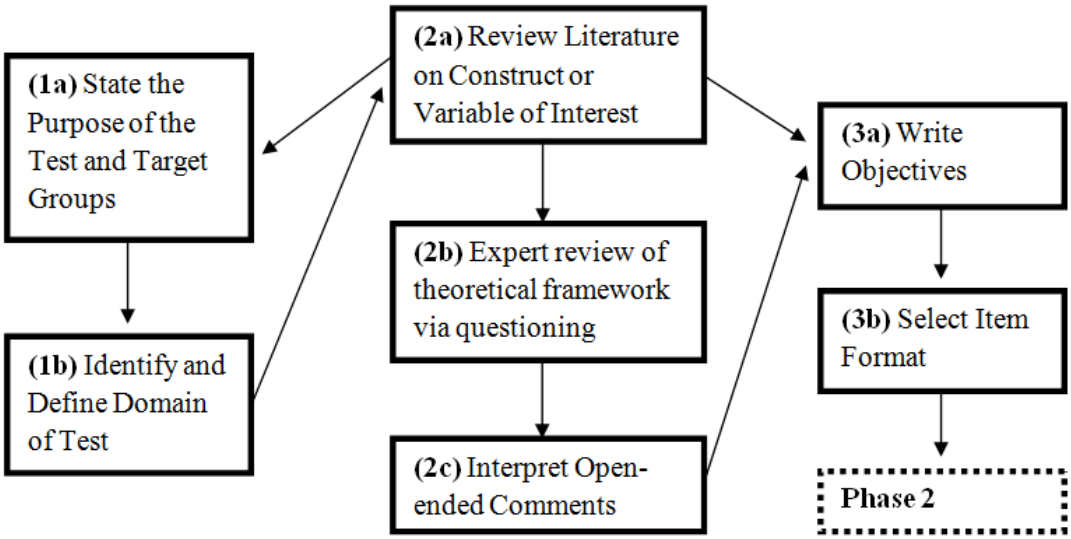

Figure 4.1. Study design of phase 1, instrument planning. Individual columns represent different steps in this phase. Phase 2 begins after step 3b is completed.

Hence, the STEMSEI creates 12 total scores for a respondent; three general scale scores measuring economic, environmental, and social sustainability separately, as well as a three subscale scores for emotional, cognitive(knowing), and cognitive (solving) engagement along each general scale. The limitations of the STEMSEI are as follows:

1. Behavioral sustainability engagement is a necessary component of the sustainability engagement framework, but is not developed/measured by the STEMSEI at this time. This portion of sustainability engagement was omitted from this study due to limitations in resources and time.
2. The STEMSEI was calibrated in this study with a sample of students almost exclusively from two universities located in the southern region of the United States. This limits the generalizability of these findings. Moreover, the item parameters estimated in this study may not be accurate reflections of

population item parameters for all post-secondary STEM students in the United States.

3. The STEMSEI should not be used as an assessment tool of any student's knowledge of sustainability.
4. The STEMSEI should not be used for evaluation of teaching.
5. The STEMSEI should not be used for the evaluation of sustainability curricula *unless* one of the curricular goals of such curriculum is to increase student sustainability engagement as previously defined. In such a case, this instrument should only be used to identify statistically significant changes in sustainability engagement and no other variable/construct. Even in such a case, the STEMSEI provides only one measure and other evaluative methods should be employed in conjunction.

Phase 1, step 2 results: Literature review and expert panel review. For results step 2a of phase 1 (literature review) (see Figure 3.1), please refer to Chapter II.

Results for steps 2b and 2c of phase 1 (expert review of theoretical framework and interpreting feedback) (see Figure 3.1) now follow. Using the qualitative study design in Figure 3.2 and the semi-structured interview protocol in Appendix C, the content experts reviewed the quantifiable definition of sustainability and the sustainability engagement framework (see Chapter II). Two rounds of interviews using the Delphi method (see Okoli & Pawlowski, 2003) were all that were needed to produce convergent agreement on all questions with all content experts.

The importance of the STEMSEI and domain feedback. Unanimously, responses to questions 1a, 1b, 1d, 1g, and 2a through 2d of the semi-structured interview protocol in Appendix C were "yes" with little additional feedback on these questions. This indicated that the content experts unanimously agreed in the first round that measuring sustainability engagement in post-secondary STEM students was important and that measures should vary across the population (i.e., that post-secondary STEM students will have varying levels of sustainability engagement). They also indicated that such a measurement would be informative to their practice as both sustainability educators and sustainability education researchers. The aforementioned data provided evidence for the relevance/utility (see Messick, 1990) of the STEMSEI. The content experts also unanimously agreed that the economic, environmental, and social domains of sustainability should be equally represented in the STEMSEI. Unanimously, the content experts felt the engagement types should also be equally represented.

The content experts unanimously responded "yes" to question 1c ("As a guiding definition for sustainability, do you think the quantifiable definition of sustainability is sufficient for this research? What additions or suggestions would you have for it?") with some additional comments. With regards to this question, the content experts specializing in engineering warned that, in practice, sustainability cannot be achieved through separate focuses on these domains; instead, sustainability must bring together all three into one unified system, as argued by Meadows (2008). However, in the event that within the population sustainability engagement functioned separately along the economic, environmental, and social domains, the content experts in engineering all felt that these three domains (economic, environmental, and social) would be sufficient to measure sustainability engagement. All content experts also thought all three domains of sustainability (economic, environmental, and social) should be equally represented in the instrument.

Question 1d ("Will all aspects of sustainability engagement be covered under cognitive and emotional engagement?" (Acknowledge the obvious limitation of omitting behavioral engagement.) received feedback related to the omission of behavioral engagement from the STEMSEI. As expected, concerns were raised about the exclusion of behavioral engagement in this study. However, the researcher explained the resource concerns for including behavior engagement, and the content experts all acceded that these concerns were well-founded for this study given its breadth and scope.

Moreover, clarification of content expert responses to question 1e ("Is the definition for the cognitive engagement construct sufficient to cover all aspects of engagement with sustainability that relies on factual information pertaining to sustainability or other cognitively driven ideas/interactions? If not, what additions would you recommend?") revealed possible sub-domains of cognitive engagement. Recall that the sustainability engagement framework (see Chapter II) provided the following examples of cognitive sustainability engagement:

- investing the mental effort to comprehending sustainability;
- knowing/comprehending complex ideas of sustainability;
- knowing/refining the skills necessary to be sustainable;
- problem solving or developing solutions for sustainability.

Four out of 6 content experts felt as though the examples of "investing the mental effort to comprehend sustainability", "knowing/comprehending complex ideas of sustainability", and "knowing/refining the skills necessary to be sustainable" might be representations of the same sub-construct of the cognitive sustainability engagement construct (i.e., knowledge of sustainability). It was suggested that these three examples were separate from "problem solving or developing solutions for sustainability" (i.e., solving for sustainability). This provided supporting evidence for subdomains along the cognitive engagement subscale, which was later found in the response data gathered in phase 3.

Synthesizing the data from round 1 to make edits, two minor changes to the theoretical framework were made and then individually verified with each content expert in a second round. First, as put forth by the content experts in engineering, the plausibility of a unified view of sustainability was considered by the other content experts, and was then unanimously supported as a possibility to consider. Second, the distinction of "knowledge for sustainability" and "problem solving or developing solutions for sustainability" was verified with all content experts. Following this, the quantifiable definition of sustainability and the sustainability engagement framework were seen as sufficiently vetted by content experts to begin forming the STEMSEI.

Phase 1, step 3 results: Objectives and item format. No results are presented for these sections. However, for reader ease, objectives developed for the STEMSEI are presented again here (see Table 4.1). For discussion on item format, please refer to the corresponding section in Chapter III.

Phase 2: Instrument Construction

The purpose of this phase of the study was first to construct an instrument blueprint for the STEMSEI as well as pool items written for the instrument. Expert review of the items was performed as well as cognitive interviews within the population. For reader ease, the study diagram for this phase is presented again here (see Figure 4.2).

Phase 2, step 4 results: Instrument blueprint and item writing. Referencing content expert feedback from step 2 of phase 1, an equal spread over all item categories was selected. The following instrument blueprint was developed for the STEMSEI (see

Table 4.2). No applicable results are presented for steps 4b and 4c of phase 2 ("train item writers" and "write pool items", respectively; see Figure 4.2).

Table 4.1
Developed Objectives for the STEMSEI

Objective	Description
1 - Sustainability Domains	This instrument will assess engagement across all domains of sustainability. See the quantifiable definition of sustainability for definitions of each domain of sustainability.
2 - Engagement Domains	This instrument will assess emotional and cognitive engagement with respect to sustainability. Emotional engagement is based on positive and negative engagement. Cognitive engagement is based on frequency or occurrence of developing solutions for sustainability. See the sustainability engagement framework for definitions of each engagement type.
3 - Differentiation	When differences exist in the population that are supported in the literature, this instrument will differentiate in the population accordingly.

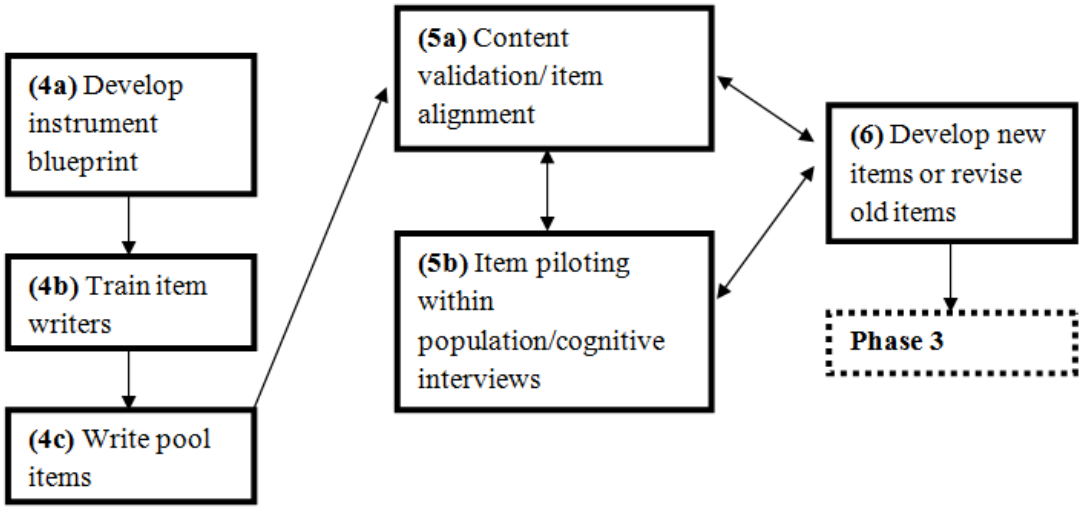

Figure 4.2. Study design of phase 2, instrument construction. Individual columns represent different steps in this phase. Phase 3 begins after step 6 is completed. The names of steps 5a and 5b were modified from the language used by Benson and Clark (1982). These changes were made to allow the names of the steps to coincide with language used in more recent methodologies corresponding to these steps.

Table 4.2
STEMSEI Blueprint

		Emotional Engagement	Cognitive Engagement	Totals
Sustainability Domains	Environmental	$16.\bar{6}\%$	$16.\bar{6}\%$	$33.\bar{3}\%$
	Economic	$16.\bar{6}\%$	$16.\bar{6}\%$	$33.\bar{3}\%$
	Social	$16.\bar{6}\%$	$16.\bar{6}\%$	$33.\bar{3}\%$
Totals		50.0%	50.0%	100.0%

Phase 2, step 5 and step 6 results: Content validation and item piloting. The purpose of steps 5 and 6 of phase 2 were to assess how well the pool items developed in step 4 might tap the constructs of interest (emotional and cognitive sustainability engagement along the economic, environmental, and social domains of sustainability). A concurrent mixed methods design (see Creswell, 2002) was employed to simultaneously gather quantitative evaluation data from content experts while gathering qualitative evaluation data from post-secondary STEM students.

Three rounds of the concurrent mixed methods design (see Figure 3.4) were completed. At the end of each round, qualitative and quantitative data was triangulated (see Creswell, Plano Clark, Gutmann, & Hanson, 2003; Erzberger & Kelle, 2003) to identify problematic items. Similarly, triangulation (see Creswell, Plano Clark, Gutmann, & Hanson, 2003; Erzberger & Kelle, 2003) was used to help identify areas for revision within an item. If revisions could not be made, the item was deleted. Both quantitative and qualitative data served to ensure the validity of any findings of the STEMSEI in phase 3.

Round 1 results. Since sustainability was being theorized as three interconnected domains (economic, environmental, and social), initial item development aimed to replicate this delineation between the sustainability domains while still preserving the unified perception of sustainability raised by the engineering content experts in phase 1. However, this was initially problematic in the first round of review by the content experts.

Since the first round of the quantitative portion of the alignment study was blind, the content experts ascribed items (1) any combination of the three domains of

sustainability and (2) one or both engagement types. For example, with respect to sustainability domains, the content experts were able to ascribe an item's content as either "economic", "environmental", or "social", or any combination of those three. The content experts ascribed multiple sustainability domains to varying degrees across the items, with approximately 55% of the items aligning with multiple sustainability domains across all content experts. The content experts in engineering ascribed more multiple sustainability domains alignments than any other discipline. On the other hand, alignments across the engagement types (emotional and cognitive) were much stronger, with only approximately 15% of the original STEMSEI items having multiple engagement types across all content experts. If a majority of the content experts agreed on the alignment of an item in both sustainability domain and engagement type, that item was considered for the item bank after triangulating (Creswell, Plano Clark, Gutmann, & Hanson, 2003; Erzberger & Kelle, 2003) cognitive interview data.

For the student cognitive interviews, most items were flagged as problematic for mostly one of two reasons. First, differences in personal social concerns or social factors between interviewees seemed to play the largest role. For example, one student interpreted an item intended to measure environmental sustainability engagement ("Environmental regulations do more harm than good.") in terms of the economic domain of sustainability based on her experience growing up in a town where coal was a major component of the local economy. Due to this interpretation, she responded "strongly agree" but in the context of the environmental regulations limiting economic growth. Though "strongly agree" could be interpreted here as low sustainability engagement in both the economic and environmental domains, this item did not accurately measure the student's environmental sustainability engagement as the item had been intended to do. No revisions were seen as possible for the item to eliminate this improper alignment in the population, so it was deleted from the item bank.

The second reason items were flagged as problematic in the cognitive interviews was due to unforeseen construct-irrelevant variance associated with some items, a strength of cognitive interviews (Conrad & Blair, 1996). For example, one item ("Humans can find ways transcend the laws of nature through science.") was omitted when a scientific misconception informed the response of an engineering graduate

student. Initially, the student debated between the responses "strongly disagree" and "disagree" for this item. His reasoning for finally choosing "disagree" was because there were certain natural limitations we (humans) could eventually exceed. In particular, this student thought engineers should eventually be able to create an engine that could go 1000 miles with one gallon of gasoline. This thought highlighted the student's misconceptions of energy transference and total potential energy within combustion engines. Across the population of post-secondary STEM students, this thought process/misconception would introduce construct-irrelevant variance in responses to this item, which would degrade the validity of the scores produced by the STEMSEI. Subsequently, this item was also deleted from the item bank.

After the first round, items that (1) showed strong alignment in both sustainability domain and engagement type and (2) showed no signs of misinterpretation within the sample were added to the item bank. Out of the original 74 items that were developed for the STEMSEI, 27 items (36%) were added to the item bank. These 27 items were analyzed for similarities to inform future item writing. Common item content with respect to sustainability domains for these 27 items ranged from economic market stability and growth, product manufacturing, recycling, renewable energy, use of non-renewable resources, environmental pollution, global population growth, and education efforts related to sustainability. In the engagement type classifications, no common themes aside from those informed by the sustainability engagement framework were identified. That is to say that each item's engagement type was reflected in the definitions of each type of engagement specified by the sustainability engagement framework.

Round 2 results. After 18 new items were written for the STEMSEI, they were once again fielded with both the content experts and post-secondary STEM students. However, the alignment study methodology in this round was slightly modified. Item alignments were no longer blind, meaning content experts rated the alignment of each item based on a predetermined sustainability domain and engagement type. Predetermination for each item was established by the researcher using round 1 results. An alignment rating for sustainability domain and engagement type was assigned separately using a 5-point Likert scale to indicate "strong alignment", "acceptable

alignment", "weak alignment", "no alignment", or that the content expert was "unable to judge" alignment of the item with respect to sustainability domain and engagement type. These methods were similar to those employed by Norcini et al. (1993) and Webb (1997). Alignment in this round was indicated by a majority of the content experts assigning a "strong alignment" or "acceptable alignment" rating to the items. By these standards, all 18 new items showed proper alignment with both their predetermined sustainability domain and engagement type.

In this round, the cognitive interview methodology used with the post-secondary STEM students was unchanged (see Appendix E). The cognitive interviews with post-secondary STEM students revealed that the construct of interest for each item seemed to be maintained across all interviewees. In particular, there were no apparent signs of construct-irrelevant variance introduced by the items. The interviewees also found most items to be more straightforward and easier to interpret than items in the first round.

Though the results for these 18 items were very encouraging, the researcher suspected that measurement efficacy could be further improved through item revision.

Round 3 results. For this round, 27 items from round 1 were revised to improve readability and interpretability using item-writing techniques employed to create the successful items from round 2. Nine items across the three sustainability domains were revised to measure emotional sustainability engagement via concern towards various sustainability related topics. Similarly, 9 revised cognitive engagement items were developed to measure self-perceived knowledge of various sustainability topics (theorized "knowledge of sustainability", a sub-domain of cognitive sustainability engagement as developed from phase 1). Finally, 9 more cognitive engagement items measuring self-perceived ability to solve various sustainability problems were also developed. These last 9 items were theorized as pertaining to the "problem solving or developing solutions for sustainability" sub-domain of cognitive sustainability engagement.

Since the new items were effectively revisions of old items whose content had already been established in prior rounds, review by content experts was seen as not necessary. Cognitive interviews were conducted with these 27 revised items to assess readability and interpretability within the post-secondary STEM student sample. Item

performance issues were not indicated in the results of the cognitive interviews. In fact, when comparing these 27 items to similar items in the previous round, the majority interviewees preferred the new items to those in round 2, commenting that the new items were easier to understand, which made responding quicker.

Finalized results. After the third round, a total of 15 items for the emotional sustainability engagement were identified, with an approximately even spread across the sustainability domains (27%, 33%, and 40% across the economic, environmental, and social domains, respectively). For the cognitive sustainability engagement, 18 items were identified that were evenly spread across the three sustainability domains. This met the goals that were outlined to proceed with the first round of piloting.

Phase 3: Instrument Evaluation

The purpose of this phase was to test the psychometric properties of the STEMSEI to assess the extent to which the instrument could produce interpretable and useable/meaningful results for sustainability educators at the post-secondary level. For reader ease, the study diagram for this phase is presented again here (see Figure 4.3).

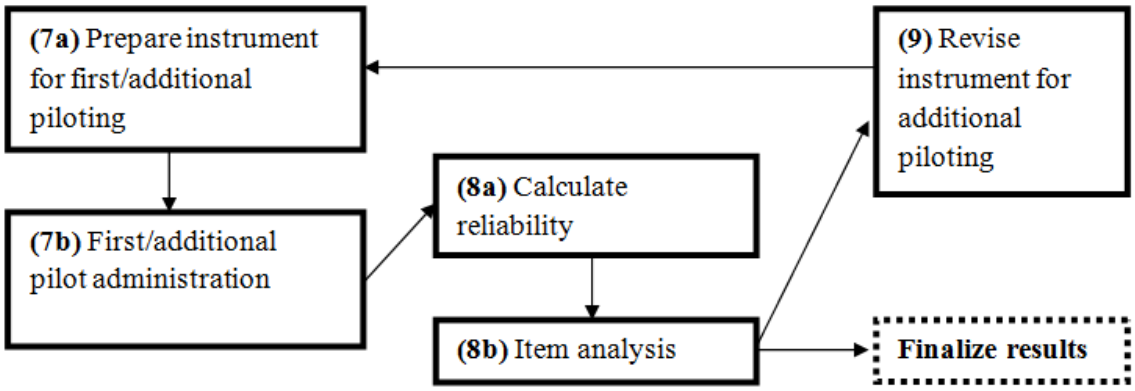

Figure 4.3. Study design of phase 3, instrument evaluation. Individual columns represent different steps in this phase.

Initially, responses to the STEMSEI from post-secondary STEM majors were gathered. Then, using the responses gathered, the quantitative properties of the STEMSEI were assessed using factor analyses, item response theory (IRT), and parametric statistical analyses. Subsequently, results from the quantitative analysis were further explored via a semi-structured interview protocol (see Appendix G). In these interviews, post-secondary STEM students were asked to contextualize differences they observed between the three sustainability domains. Results from both the quantitative

and qualitative portions of this phase informed the second round of instrument piloting in two ways. First, since the instrument was not functioning in a unidimensional manner, it informed which areas of the instrument needed additional items and possible reasons for these needs. Second, it informed which scores and sub-scores produced by the STEMSEI along different sustainability domains were contextually reflected within this sample. This phase was completed after one round of both quantitative and qualitative portions of the design and one final piloting round.

Phase 3, step 7, round 1 results: Preparing for instrument piloting and pilot administration. Three post-secondary institutions (PSI) participated in phase 3 of the study. At PSIA, there were N_1 = 488 total responses out of 21,295 study invitations sent out, resulting in a response rate of 2.3%. There were N_2 = 245 and N_3 = 37 responses from PSIB and PSIC, respectively. Due to the sampling methods employed at the second and third institutions, response rates had to be estimated and were 3.0% and 7.4%, respectively. The overall missing data rate was 7.3%. All missing data were treated as missing not at random. Across all items, all response categories were used to varying degrees.

Phase 3, step 8a, round 1 results: Calculate reliability. Across all STEMSEI items in the complete sample, Cronbach's alpha was $\alpha = .861$. Per Kline's (2000) recommendations, this was sufficient to proceed. Omitting an item decreased reliability for all but two items. For the other two, reliability increased no more than .01. Hence, no items were dropped.

Phase 3, step 8b, round 1 results: Item analysis. This step of phase 3 evaluated how the STEMSEI items performed in the first round of piloting. The following results first detail how item analysis transitioned from factor analysis results to item response theory results (i.e., quantitative piloting). Results of the semi-structured interviews (i.e., qualitative piloting) concerning results obtained in the quantitative piloting follow.

Round 1 factor analysis results with development sample. Results for the factor analyses with the development sample are first provided with the factor analysis results of the validation sample following.

Three different approaches were taken to the factor analyses with the development sample. First, as suggested by the content experts, an exploratory factor

analysis (EFA) on all 33 items was attempted. Two separate EFA's on the emotional and cognitive engagement items were also estimated. Finally, three separate EFA's on each sustainability domain were estimated. All EFA's employed an oblimin rotation. Results for confirmatory bifactor model solutions then follow for EFA solutions with strong fit within the development sample.

The EFA with all 33 items estimated eight eigenvalues greater than 1. Consequently, EFA's ranging from 1-factor up to 8-factors were extracted and compared for fit. This resulted in a comparative fit index (CFI) values ranging from .561 to .913 and root mean square error of approximation (RMSEA) values ranging from .066 to .134. Fit indices for the 1-factor solution up to the 4-factor solution were not within tolerable range. While fit indices for the 5-factor solution to the 8-factor solution were within range, none of the rotated factor structures were interpretable for these models. Hence, a unidimensional view of sustainability was not seen as possible.

Results fared better for the two EFA's conducted separately on the emotional engagement and cognitive engagement items. The EFA on all 15 emotional engagement items resulted in 4 eigenvalues greater than 1, as did the EFA on the 18 cognitive engagement items. EFA's ranging from 1-factor to 4-factors were extracted and interpreted for both the emotional engagement and cognitive engagement items (see Table 4.3). The statistically significant results of the chi-square tests are not surprising

Table 4.3

EFA's for Emotional Engagement and Cognitive Engagement Items

	Dimensions	Cronbach's Alpha	Chi-Square Test			CFI	RMSEA
			χ^2	*df*	*p*		
Emotional Engagement	1-factor	α = .805	678.177	90	< .0001	.808	.131
	2-factors		249.820	76	< .0001	.943	.059
	3-factors		150.885	63	< .0001	.971	.061
	4-factors		112.339	51	< .0001	.980	.056
Cognitive Engagement	1-factor	α = .840	3206.502	135	< .0001	.722	.183
	2-factors		1984.443	118	< .0001	.831	.153
	3-factors		736.811	102	< .0001	.942	.096
	4-factors		360.541	87	< .0001	.975	.068

Note. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

since this statistic is overly sensitive to larger samples (Meyers, Gamst, & Guarino, 2006). While the CFI and RMSEA were within acceptable range for many of these extractions, the associated factor structures were mostly not interpretable. Only the 2-factor extraction was interpretable in the emotional engagement extraction. In this extraction, the economic items tended to load onto the first factor while the environmental and social items both loaded onto the second factor. This extraction was initially seen as pursuable, but was outperformed by the EFA results to follow.

When exploring each sustainability domain separately, EFA results were very strong for the economic and environmental domains, while EFA extractions in the social domain were less successful (see Table 4.4). All fit statistics indicated model-fit for the 3-factor economic scale, while the RMSEA estimate for the environmental scale was large and the chi-square test was statistically significant (see Table 4.4). This meant the model for the environmental scale should be interpreted with caution. There were three eigenvalues greater than 1 for the economic and environmental domain extractions, while there were four for the social domain. In order, the first three eigenvalues for the 3-factor extractions of the economic and environmental domains were $\lambda_1 = 4.1$, $\lambda_2 = 1.7$, $\lambda_3 = 1.4$ and $\lambda_1 = 5.2$, $\lambda_2 = 1.9$, $\lambda_3 = 1.4$. A clear factor structure was identified from the 3-factor extractions for both the economic and environmental domains (see Table 4.5), essentially separating the items along the engagement types for which they were written. Moreover, using recommendations from Matsunaga (2010), no cross-loading was identified for any

Table 4.4

EFA's for Economic, Environmental, and Social Items

		Cronbach's Alpha	Chi-Square Test			CFI	RMSEA
			χ^2	*df*	*p*		
Economic Scale	1-factor	α = .794	692.699	35	< .0001	.844	.222
	2-factor		403.223	26	< .0001	.910	.195
	3-factor		27.830	18	.0647	.998	.038
Environmental Scale	1-factor	α = .846	831.017	54	< .0001	.776	.195
	2-factor		504.717	43	< .0001	.867	.168
	3-factor		126.587	33	< .0001	.973	.087
Social Scale	1-factor	α = .662	221.585	44	< .0001	.739	.103
	2-factor		133.077	34	< .0001	.855	.087
	3-factor		81.214	25	< .0001	.918	.077
	4-factor		35.903	17	.0047	.972	.054

Note. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

Table 4.5
EFA Factor Structures for Economic and Environmental Items

	Item	Factor Patterns 1	2	3	Engagement Type
Economic Scale	1	**0.676**	0.074	-0.041	Emotional
	2	**0.759**	-0.031	-0.049	Emotional
	3	**0.730**	0.001	0.093	Emotional
	4	**0.335**	0.221	0.144	Emotional
	5	-0.012	**0.883**	0.040	Cognitive (Know)
	6	-0.036	**0.981**	-0.027	Cognitive (Know)
	7	0.080	**0.853**	0.005	Cognitive (Know)
	8	0.071	-0.055	**0.819**	Cognitive (Solution)
	9	-0.045	0.024	**0.823**	Cognitive (Solution)
	10	-0.062	0.074	**0.646**	Cognitive (Solution)
Environmental Scale	1	**0.814**	0.058	-0.057	Emotional
	2	**0.912**	-0.024	0.004	Emotional
	3	**0.791**	-0.060	0.089	Emotional
	4	**0.476**	-0.052	-0.036	Emotional
	5	**0.538**	0.133	0.186	Emotional
	6	**0.731**	0.043	-0.045	Emotional
	7	-0.071	**0.987**	-0.009	Cognitive (Know)
	8	0.254	**0.630**	-0.024	Cognitive (Know)
	9	0.077	**0.684**	0.117	Cognitive (Know)
	10	-0.057	0.069	**0.820**	Cognitive (Solution)
	11	0.067	0.004	**0.799**	Cognitive (Solution)
	12	0.005	-0.041	**0.871**	Cognitive (Solution)

of the items. On the other hand, no determinable factor structure for any of the social domain extractions could be identified. The low reliability observed in the social items (see Table 4.4) likely contributed to this result.

Since the EFA's along the sustainability domains produced more interpretable factor structures stable results with the best fit statistics, confirmatory bifactor models for the economic and environmental domains using similar factor structures were also estimated (see Table 4.6) for comparison to the standard EFA models. Again, normally an exploratory bifactor analysis would be recommended here, but software for such an analysis was not available to the researcher but does exist. These results indicated that the bifactor model more appropriately fit the environmental scale than the corresponding 3-factor EFA structure. This increase in fit was interpreted to be due in part to the degree of unidimensionality present in the factor structures. Recall that a bifactor model

Table 4.6
Corresponding Bifactor Model Extractions for Economic and Environmental Items

		Chi-Square Test			CFI	RMSEA
		χ^2	*df*	*p*		
Economic Scale	Bifactor Model (3-factors)	52.035	25	.0012	.994	.053
Environmental Scale	Bifactor Model (3-factors)	126.906	42	<.0001	.976	.073

Note. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

accounts for unidimensionality while simultaneously accounting for the multidimensional factor structure (Reise, Moore, & Haviland, 2013); standard EFA models cannot achieve this dual perspective. In contrast, model fit with the economic scale lessened to a slight degree with the bifactor model. However, both were considered for possible IRT models since model-fit was acceptable for both models across all fit statistics.

Round 1 factor analysis results with validation sample. Results from the factor analyses with the development sample indicated that the economic and environmental scales were 3-dimensional. The three subscales for these scales were along the emotional engagement items, cognitive (know) items, and cognitive (solution) items (see Table 4.5). Using a similar factor structure, confirmatory factor analysis results on the economic and environmental scale responses of the validation sample indicates strong model-fit across two of three of the fit statistics (see Table 4.7). Again, the chi-square test for model fit is sensitive to large samples (Meyers, Gamst, & Guarino, 2006), so the statistically significant chi-square test of model fit result was interpreted with caution. Essentially,

Table 4.7
Confirmatory Factor Analysis and Confirmatory Bifactor Analysis Results for Validation Sample

		Chi-Square Test			CFI	RMSEA
		χ^2	*df*	*p*		
Economic Scale	CFA 3-factor	200.341	32	< .0001	.960	.118
	Confirmatory Bifactor	59.507	25	.0001	.992	.062
Environmental Scale	CFA 3-factor	180.406	51	< .0001	.963	.082
	Confirmatory Bifactor	130.412	42	< .0001	.971	.077

Note. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

these results indicated that a general factor structure was repeated across both the development and validation samples, providing evidence for the generalizability of the instrument. Also of note is that the confirmatory bifactor model fit both scales better than a traditional 3-dimensional CFA, evidenced by CFI values closer to 1 and RMSEA values closer to 0 (see Table 4.7). Since there were no interpretable EFA results for the social scale, neither a CFA or a confirmatory bifactor analysis were performed. This may have been due to the low reliability estimate for the social items in this round.

Round 1 item response theory results. Results of the factor analysis results indicated stronger model-fit along sustainability domains. Hence, item response theory (IRT) models considered here are presented in a similar fashion (i.e., as economic, environmental, and social scales). Since a clear factor structure could not be identified for the social scale, an IRT analysis along the social domain was not possible in the first round. However, for both the economic and environmental domains, only multidimensional item response theory (MIRT) models should be considered since the factor structure of these two domains were clearly multidimensional (de Ayala, 2009). Multidimensional versions of the generalized partial credit model (GPCM; see Reckase, 2009) and the graded response model (GRM; see Reckase, 2009) were considered as well as the bifactor IRT model. The former two models were considered due to the strong fit of the EFA and CFA models in the development and validation samples, respectively. The bifactor IRT model was also considered since the confirmatory bifactor analyses also indicated strong model-fit.

Initial models calculated for the economic and environmental scales using the multidimensional GPCM, multidimensional GRM, and the bifactor IRT model indicated some misfit of the model at the item-level. However, rather than detail these item omissions for all models, evidence of the best model is presented first. A detailing of the item omission process for that particular model follows after the next paragraph.

With both the economic and environmental scales, use of the AIC Difference statistic (see Burnham & Anderson, 2002) showed that the fit of the multidimensional GPCM was considerably less than both the multidimensional GRM and the bifactor model (see Table 4.8). Due to this, the GPCM was not further considered. On the other hand, the AIC Difference statistic between the multidimensional GRM and the bifactor

Table 4.8
General Partial Credit Model (GPCM), Graded Response Model (GRM), and bifactor IRT models Akaike information criterion (AIC) statistics and AIC Difference statistics

	Model	AIC	AIC Difference
Economic Scale	3-factor GPCM	6346.50	
	3-factor GRM	6340.81	5.69
	3-factor GPCM	6346.50	
	Bifactor Model	6342.39	4.11
	3-factor GRM	6340.81	
	Bifactor Model	6342.39	1.58
Environmental Scale	3-factor GPCM	7272.27	
	3-factor GRM	7250.36	21.91
	3-factor GPCM	7272.27	
	Bifactor Model	7235.76	36.51
	3-factor GRM	7250.36	
	Bifactor Model	7235.76	14.6

model was less than 2 for the economic scale while it was greater than 10 for the environmental scale (see Table 4.8). This meant there was essentially no empirical support of the multidimensional GRM over the bifactor model for the environmental scale. Moreover, the theoretical value of the bifactor model far outweighs that of the multidimensional GRM since it can produce a general scale score as well as subscale scores (Reise et al., 2013). In comparison, the multidimensional GRM will only produce a score along each subscale and not produce a general score for each participant. Considering this and the lack of empirical support for the multidimensional GRM in the environmental scale, the multidimensional GRM was also omitted from further consideration.

Since the bifactor IRT model was selected as the final model for the STEMSEI, the item omission processes for this model on both the economic and environmental scales are now presented. The initial bifactor IRT model along the economic items estimated item slopes (*a*) greater than 3 for three items ($3.15 \leq \alpha \leq 4.15$). To control for these large slopes, prior parameters using a normal distribution ($\mu = 1.7$, $\sigma = 0.5$) were then implemented for item slope parameter estimates. The resultant bifactor IRT model

then estimated a negative item slope for one of the items ("I am interested in using my major to help businesses operate efficiently."). This meant the model was interpreting lower levels of endorsement to indicate higher levels of sustainability engagement; however, this is counter to the item's intent with respect to the construct of interest. This item was removed, and the model recalibrated. Another item ("How much do you know about making production manufacturing more efficient for businesses?") was then omitted due to a large item slope parameter (a = 4.23) even after using prior parameter settings. A bifactor IRT model along the economic domain was then found and fit statistics for the model were all within range (see Table 4.9).

Following a similar process, a bifactor IRT model for the environmental domain items was calculated (see Table 4.10). Two items were omitted sequentially from the

Table 4.9
Initial Bifactor IRT Model Parameters for Economic Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistic			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p*	*B.H.p**
1	1.14 (0.20)	1.35 (0.24)	0 (-)	0 (-)	3.28 (0.29)	0.49 (0.16)	-2.48 (0.24)	56.77	36	0.015	.040
2	1.02 (0.23)	2.11 (0.48)	0 (-)	0 (-)	4.94 (0.74)	1.96 (0.34)	-1.40 (0.25)	43.02	35	0.165	.220
3	1.37 (0.24)	1.58 (0.29)	0 (-)	0 (-)	4.07 (0.39)	1.02 (0.19)	-2.28 (0.26)	47.63	35	0.075	.150
4	2.23 (0.32)	0 (-)	2.51 (0.32)	0 (-)	1.73 (0.27)	-3.40 (0.30)	-6.74 (0.49)	25.32	31	0.754	.754
5	2.27 (0.30)	0 (-)	2.51 (0.32)	0 (-)	2.19 (0.29)	-3.10 (0.30)	-6.70 (0.46)	36.6	32	0.263	.300
6	1.65 (0.28)	0 (-)	0 (-)	1.94 (0.35)	3.23 (0.36)	0.40 (0.20)	-2.04 (0.29)	42.28	33	0.129	.206
7	1.52 (0.26)	0 (-)	0 (-)	2.00 (0.39)	3.48 (0.39)	0.69 (0.21)	-1.99 (0.28)	54.58	32	0.008	.040
8	1.00 (0.18)	0 (-)	0 (-)	1.37 (0.24)	1.17 (0.18)	-0.71 (0.17)	-2.33 (0.23)	55.63	35	0.014	.040

Note. Values in ( ) are parameter estimate standard errors. a = item slope; b = item location. These values are for a logistic scale.
*B.H.p = Benjamini-Hochberg correction values

Table 4.10
Initial Bifactor IRT Model Parameters for Environmental Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistic			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p*	*B.H.p*
1	1.79 (0.24)	1.97 (0.28)	0 (-)	0 (-)	6.67 (0.63)	3.25 (0.35)	0.01 (0.23)	36.39	32	0.271	.490
2	2.14 (0.27)	2.43 (0.29)	0 (-)	0 (-)	6.39 (0.59)	3.62 (0.38)	-0.06 (0.26)	36.77	34	0.341	.490
3	1.62 (0.21)	1.64 (0.23)	0 (-)	0 (-)	6.08 (0.56)	3.31 (0.32)	0.07 (0.20)	24.29	29	0.715	.715
4	1.89 (0.25)	0.76 (0.20)	0 (-)	0 (-)	4.45 (0.40)	2.24 (0.26)	-0.89 (0.20)	51.65	37	0.055	.275
5	1.44 (0.19)	1.45 (0.24)	0 (-)	0 (-)	3.73 (0.34)	1.39 (0.22)	-2.65 (0.25)	35.57	40	0.671	.715
6	1.68 (0.22)	0 (-)	1.26 (0.29)	0 (-)	5.78 (0.58)	1.82 (0.24)	-1.62 (0.22)	60.37	36	0.007	.070
7	1.66 (0.23)	0 (-)	1.26 (0.29)	0 (-)	3.94 (0.44)	0.23 (0.19)	-3.12 (0.31)	51.06	42	0.159	.490
8	1.68 (0.26)	0 (-)	0 (-)	2.09 (0.29)	3.35 (0.35)	0.81 (0.23)	-1.47 (0.25)	50.6	44	0.229	.490
9	1.84 (0.27)	0 (-)	0 (-)	1.79 (0.26)	3.49 (0.34)	1.37 (0.24)	-1.20 (0.23)	45.09	42	0.343	.490
10	1.75 (0.27)	0 (-)	0 (-)	2.12 (0.30)	2.88 (0.32)	0.55 (0.23)	-1.90 (0.28)	43.56	44	0.491	.614

Note. Values in ( ) are parameter estimate standard errors. *a* = item slope; *b* = item location. These values are for a logistic scale.
*B.H.p = Benjamini-Hochberg correction values

environmental domain analyses ("I am not worried about finding renewable sources of electricity," and "how much do you know about current consumption rates of fossil fuels") because of item slope parameters (a = 0.29, 4.01, respectively) outside the recommended bounds of .5 to 3 (see Baker, 2001). The standardized LD χ^2 statistics and the S-χ^2 item level diagnostic statistics were all within tolerable bounds for the final environmental domain bifactor IRT model (see Table 4.10).

Unidimensional DIF analyses along the economic scale revealed no DIF (see Table 4.11). However, initial unidimensional DIF analyses of the environmental scale

Table 4.11

Phase 3, Round 1: Differential Item Functioning Tests for Economic Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p**	X^2_a	*df*	*p*	*B.H. p**	$X^2_{c\|a}$	*df*	*p*	*B.H. p**
Race	1	11.9	4	0.0182	0.1820	0.1	1	0.7458	-	11.8	3	0.0082	-
	2	5.3	4	0.2605	0.6977	0.8	1	0.3619	-	4.5	3	0.2178	-
	3	9.4	4	0.0512	0.2560	1.1	1	0.2949	-	8.3	3	0.0397	-
	4	3.9	4	0.4186	0.6977	0.1	1	0.7620	-	3.8	3	0.2823	-
	5	4.5	4	0.3421	0.6977	0	1	0.8294	-	4.5	3	0.2165	-
	6	3.3	4	0.5121	0.7316	0.5	1	0.4926	-	2.8	3	0.4222	-
	7	1.6	4	0.8005	0.8793	0	1	0.9440	-	1.6	3	0.6502	-
	8	1.2	4	0.8793	0.8793	0.2	1	0.6793	-	1	3	0.7960	-
	9	4.4	4	0.3590	0.6977	0.9	1	0.3393	-	3.5	3	0.3274	-
	10	1.3	4	0.8588	0.8793	0.8	1	0.3681	-	0.5	3	0.9180	-
Present Religion	1	6.5	4	0.1660	0.8053	1.6	1	0.2100	-	4.9	3	0.1808	-
	2	12.4	4	0.0143	0.1430	5.4	1	0.0197	-	7	3	0.0716	-
	3	2.6	4	0.6271	0.8166	0.4	1	0.5147	-	2.2	3	0.5373	-
	4	3.7	4	0.4535	0.8070	0	1	0.8528	-	3.6	3	0.3047	-
	5	2.5	4	0.6533	0.8166	1.5	1	0.2247	-	1	3	0.8075	-
	6	3.5	4	0.4842	0.8070	0.1	1	0.7234	-	3.3	3	0.3433	-
	7	0.9	4	0.9235	0.9235	0.1	1	0.7190	-	0.8	3	0.8548	-
	8	2	4	0.7404	0.8227	0.1	1	0.7336	-	1.9	3	0.6023	-
	9	3.7	4	0.4496	0.8070	0	1	0.9808	-	3.7	3	0.2973	-
	10	5.5	4	0.2416	0.8053	0.4	1	0.5453	-	5.1	3	0.1623	-
Gender	1	14.1	4	0.0071	0.0710	3.9	1	0.0492	-	10.2	3	0.0169	-
	2	3.4	4	0.4926	0.8146	0.5	1	0.4902	-	2.9	3	0.4030	-
	3	5.7	4	0.2250	0.5625	1.7	1	0.1889	-	4	3	0.2673	-
	4	2	4	0.7331	0.8146	1.1	1	0.2969	-	0.9	3	0.8195	-
	5	2.4	4	0.6555	0.8146	1.1	1	0.2926	-	1.3	3	0.7220	-
	6	1	4	0.9166	0.9166	0	1	0.8325	-	0.9	3	0.8231	-
	7	2.1	4	0.7266	0.8146	1.2	1	0.2818	-	0.9	3	0.8281	-
	8	3.1	4	0.5501	0.8146	0	1	0.8565	-	3	3	0.3898	-
	9	6	4	0.1959	0.5625	0.1	1	0.8002	-	6	3	0.1127	-
	10	9.9	4	0.0420	0.2100	1.1	1	0.3044	-	8.8	3	0.0313	-

Table 4.11 (cont.)
Phase 3, Round 1: Differential Item Functioning Tests for Economic Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p**	X^2_a	*df*	*p*	*B.H. p**	$X^2_{c\|a}$	*df*	*p*	*B.H. p**
Classification	1	5.2	4	0.2653	0.4422	3	1	0.0814	-	2.2	3	0.5323	-
	2	6.4	4	0.1711	0.3422	2.3	1	0.1267	-	4.1	3	0.2572	-
	3	3.3	4	0.5065	0.7236	0.2	1	0.6529	-	3.1	3	0.3747	-
	4	0.3	4	0.9880	0.9880	0.2	1	0.6227	-	0.1	3	0.9936	-
	5	14.6	4	0.0055	0.0550	3.7	1	0.0541	-	10.9	3	0.0120	-
	6	1.1	4	0.8960	0.9880	0.5	1	0.4853	-	0.6	3	0.8961	-
	7	1.9	4	0.7622	0.9528	0.5	1	0.4654	-	1.3	3	0.7238	-
	8	7.8	4	0.0980	0.3267	0	1	0.8518	-	7.8	3	0.0505	-
	9	6.8	4	0.1439	0.3422	0.1	1	0.7895	-	6.8	3	0.0793	-
	10	12.3	4	0.0151	0.0755	2.3	1	0.1334	-	10.1	3	0.0181	-

Note. Cells with a "-" entry indicate that the statistic was not interpreted since the omnibus statistic was not statistically significant.
*B.H.p = Benjamini-Hochberg correction value

revealed possible DIF in several items along gender and race (see Table 4.12). For gender, secondary unidimensional DIF analyses revealed that DIF was possible in items 7 through 10 of the environmental scale (see Table 4.13). However, closer inspection of the unidimensional models calculated in the gender DIF analyses in the environmental scale showed local dependence in items 7 through 10 ($13.3 \leq$ LD-χ^2 $(x) \leq 25.7$). Normally, these items would have been omitted from the unidimensional model. Due to this, these were seen as false positives of DIF and these items were not removed. Similarly, item 7 displayed DIF with respect to race in both the initial and secondary DIF analyses, but local dependence was also observed ($11.0 \leq$ LD-χ^2 $(x) \leq 36.3$) in the two models associated with this DIF (see Table 4.13). Due to this, the DIF results for this item were considered a false positive as well, and the item was not removed.

Marginal reliability and the standard error of estimate (SEE) for person location scores on the general factor of the economic and environmental scales indicated strongest measurement accuracy on the interval [-2, 2] (see Table 4.14). Though the SEE values for all subscales were near that of the general factor, reliability for these scales was not considered since all subscales consisted of only 2 to 4 items, far below Bohrnstedt's (2010) recommendation of at least 7 to 10 items per subscale (p. 375).

Table 4.12
Phase 3, Round 1: Initial Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p**	X^2_a	*df*	*p*	*B.H. p**	$X^2_{c\|a}$	*df*	*p*	*B.H. p**
Race	1	15.6	4	**0.0035**	**0.0117**	9.8	1	**0.0017**	**0.0026**	5.8	3	0.1203	0.1203
	2	19.9	4	**0.0005**	**0.0050**	14.3	1	**0.0002**	**0.0001**	5.5	3	0.1365	0.1365
	3	5.9	4	0.2070	0.2957	3.9	1	-	-	2	3	-	-
	4	7.2	4	0.1247	0.2494	0	1	-	-	7.2	3	-	-
	5	10.7	4	0.0300	0.0750	5.9	1	-	-	4.8	3	-	-
	6	0.8	4	0.9429	0.9429	0.6	1	-	-	0.2	3	-	-
	7	16.4	4	**0.0025**	**0.0117**	4.5	1	**0.0330**	**0.0330**	11.9	3	0.0078	0.0078
	8	4.2	4	0.3828	0.4253	3.2	1	-	-	1	3	-	-
	9	6	4	0.2038	0.2957	4.6	1	-	-	1.3	3	-	-
	10	5	4	0.2906	0.3633	3.7	1	-	-	1.3	3	-	-
Present Religion	1	11.4	4	0.0221	0.2210	0.2	1	-	-	11.3	3	-	-
	2	5.5	4	0.2387	0.3978	4.6	1	-	-	1	3	-	-
	3	3	4	0.5510	0.6888	2.3	1	-	-	0.8	3	-	-
	4	1.4	4	0.8372	0.8372	0.5	1	-	-	0.9	3	-	-
	5	2.4	4	0.6622	0.7358	0.8	1	-	-	1.6	3	-	-
	6	6.4	4	0.1713	0.3426	4.4	1	-	-	2	3	-	-
	7	3.9	4	0.4275	0.6107	3.1	1	-	-	0.7	3	-	-
	8	6.6	4	0.1612	0.3426	0	1	-	-	6.5	3	-	-
	9	9.1	4	0.0586	0.2930	0.6	1	-	-	8.5	3	-	-
	10	7	4	0.1359	0.3426	0.1	1	-	-	6.9	3	-	-
Gender	1	10.4	4	0.0336	0.0672	0.3	1	-	-	10.1	3	-	-
	2	5.7	4	0.2270	0.3783	0	1	-	-	5.7	3	-	-
	3	2.4	4	0.6674	0.6674	0	1	-	-	2.4	3	-	-
	4	3.9	4	0.4216	0.5270	0.8	1	-	-	3.1	3	-	-
	5	3.4	4	0.5008	0.5564	0.6	1	-	-	2.8	3	-	-
	6	4.8	4	0.3113	0.4447	3.3	1	-	-	1.5	3	-	-
	7	30.8	4	**0.0001**	**< 0.0001**	2.2	1	0.1370	0.1370	28.6	3	**0.0001**	**0.0004**
	8	24.4	4	**0.0001**	**< 0.0001**	11	1	**0.0009**	**0.0036**	13.5	3	**0.0037**	**0.0049**
	9	21.6	4	**0.0002**	**0.0007**	5.8	1	**0.0161**	**0.0215**	15.8	3	**0.0012**	**0.0024**
	10	17	4	**0.0019**	**0.0048**	6.6	1	**0.0102**	**0.0204**	10.4	3	**0.0152**	**0.0152**

Table 4.12 (cont.)

Phase 3, Round 1: Initial Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p**	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Classification	1	4.2	4	0.3809	0.7029	3.4	1	-	-	0.8	3	-	-
	2	0.7	4	0.9488	0.9899	0.3	1	-	-	0.4	3	-	-
	3	3	4	0.5623	0.7029	0	1	-	-	3	3	-	-
	4	8.9	4	0.0640	0.6400	0	1	-	-	8.9	3	-	-
	5	3.9	4	0.4157	0.7029	0.4	1	-	-	3.6	3	-	-
	6	4.4	4	0.3573	0.7029	0.7	1	-	-	3.7	3	-	-
	7	0.3	4	0.9899	0.9899	0.2	1	-	-	0.1	3	-	-
	8	3.7	4	0.4476	0.7029	1.3	1	-	-	2.4	3	-	-
	9	3.1	4	0.5375	0.7029	1.6	1	-	-	1.6	3	-	-
	10	3.9	4	0.4271	0.7029	0.1	1	-	-	3.8	3	-	-

Note. Cells with a "-" entry indicate that the statistic was not interpreted since the omnibus statistic was not statistically significant.
*B.H.p = Benjamini-Hochberg correction value

Table 4.13

Phase 3, Round 1: Follow-up Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p**	X^2_a	*df*	*p*	*B.H. p**	$X^2_{c\|a}$	*df*	*p*	*B.H. p**
Race	1	**12.8**	**4**	**0.0125**	**0.0375**	4.2	1	0.0404	0.0536	8.6	3	0.0357	0.0536
	2	9.9	4	0.0417	0.0536	5.6	1	-	-	4.4	3	-	-
	7	**15.4**	**4**	**0.0040**	**0.0180**	1.4	1	0.2388	0.2388	14	3	**0.0029**	**0.0180**
Gender	7	34	4	**0.0001**	**0.0003**	2.6	1	0.1086	0.1086	31.4	3	**0.0001**	**0.0003**
	8	28.7	4	**0.0001**	**0.0003**	13.3	1	**0.0003**	**0.0006**	15.4	3	**0.0015**	**0.0023**
	9	25.9	4	**0.0001**	**0.0003**	7.6	1	**0.0058**	**0.0065**	18.3	3	**0.0004**	**0.0007**
	10	21.3	4	**0.0003**	**0.0006**	8.9	1	**0.0029**	**0.0039**	12.4	3	**0.0060**	**0.0065**

Note. Cells with a "-" entry indicate that the statistic was not interpreted since the omnibus statistic was not statistically significant.
*B.H.p = Benjamini-Hochberg correction value

Validation assessment results. A bifactor IRT model (i.e., a new sets of item parameters) was estimated using the validation sample for the economic and environmental scale. Corresponding item parameter estimates between each model (the development sample model and the validation sample model) were correlated using Pearson's *r*. For the two parameterizations of the economic scale, statistically significant positive correlations were found between the item slope parameters, $r = .764$, $N = 16$, $p = .001$, and the item location parameters, $r = .995$, $N = 24$, $p < .001$. Similarly for the

Table 4.14
Standard Error of Estimate (SEE) and Marginal Reliability Estimates for STEMSEI Scales

		SEE	Marginal Reliability and Corresponding Interval		
			[-3, 3]	[-2, 2]	[-1, 1]
Economic Scale	General Factor	.67	.76	.78	.78
	Emotional Subscale	.66	.60	.65	.69
	Cognitive (Knowing) Subscale	.70	.69	.71	.72
	Cognitive (Solving) Subscale	.67	.68	.70	.71
Environmental Scale	General Factor	.57	.84	.85	.85
	Emotional Subscale	.68	.59	.62	.64
	Cognitive (Knowing) Subscale	.82	.72	.75	.77
	Cognitive (Solving) Subscale	.64	.67	.70	.72

environmental scale, statistically significant positive correlations were found between the item slope parameters, $r = .918$, $N = 21$, $p < .001$, and the item location parameters, $r = .991$, $N = 30$, $p < .001$. For both models, this indicates that, under a linear transformation, the item parameters for the development sample are essentially equivalent to the item parameters for the validation sample. This provides more evidence for the generalizability of the instrument.

Next, the two sets of item parameters estimated above were used to score the validation sample; once using item parameters estimated from the development sample, another using item parameters estimated from the validation sample. The two sets of scores for each participant in the validation sample were then correlated. Statistically significant correlations near 1 on all corresponding factors (i.e., general scale to general scale, emotional subscale to emotional subscale, etc.) were observed across all corresponding factor pairs (see Table 4.15), which indicates stability between the person location estimates produced by the two models. That is to say, person location estimates

Table 4.15

Correlations of Person Scores (θ) Using Two Item Parameter Sets Across the Validation Sample and Development Sample

Environmental Scale		θ Estimated Via Validation Sample Item Parameters			
		General Factor Scores	Emotional Subscale Scores	Cognitive (Knowing) Subscale Scores	Cognitive (Solving) Subscale Scores
θ Estimated Via Development sample Item Parameters	General Factor Scores	r = .992 N = 356 p < .001	-	-	-
	Emotional Sub factor Scores	r = .354 N = 356 p < .001	r = .995 N = 356 p < .001	-	-
	Cognitive (Knowing) Subscale Scores	r = .353 N = 324 p < .001	r = -.243 N = 324 p < .001	r = .995 N = 324 p < .001	-
	Cognitive (Solving) Subscale Scores	r = .332 N = 312 p < .001	r = -.324 N = 312 p < .001	r = -.193 N = 312 p = .001	r = .987 N = 311 p < .001

Economic Scale		θ Estimated Via Validation Sample Item Parameters			
		General Factor Scores	Emotional Subscale Scores	Cognitive (Knowing) Subscale Scores	Cognitive (Solving) Subscale Scores
θ Estimated Via Development sample Item Parameters	General Factor Scores	r = .995, N = 358 p < .001	-	-	-
	Emotional Subscale Scores	r = .285 N = 357 p < .001	r = .991 N = 357 p < .001	-	-
	Cognitive (Knowing) Subscale Scores	r = .509 N = 323 p < .001	r = -.186 N = 322 p < .001	r = .995 N = 323 p < .001	-
	Cognitive (Solving) Subscale Scores	r = .341 N = 311 p < .001	r = -.230 N = 311 p < .001	r = -.235 N = 311 p < .001	r = .976 N = 311 p < .001

are invariant across the two models, indicating more evidence for the generalizability.

Next, for all items across the economic and environmental models, the RMSD on the interval [-3, 3] ranged from .04 to .11 and .03 to .09, respectively. The average RMSD value (i.e., the average across all items) was .06 for both models. This means that the probabilistic estimation option response functions for each item differed, on average, by 6%, meaning there is small difference between the models. This measure also provides evidence for the generalizability of the instrument across different samples.

Now, recall that IRT models always fit the sample in which they are calibrated better (de Ayala, 2009). MODFIT 3.0 (Stark, 2008) was used to assess how well model-fit could be achieved for responses in validation sample compared to the option response curves generated by the development sample. For the economic scale, empirical fit for most response options on all items fared well. However, there was some degree of model-misfit for one item ("To what extent could your major help businesses stay profitable?") with the lower response options ("To no extent" and "To a small extent"). This item also had the largest RMSD value for the economic scale. Both of these misfitting empirical curves were parallel to their corresponding option response function (see Figure 4.4), indicating that within this sample, the corresponding item threshold between these two response options would be estimated to be slightly higher on the latent trait continuum. This essentially meant that, for this item in the validation sample, these two options were easier to endorse when compared to the development sample. In contrast, the environmental scale empirically fit all items on all response options well

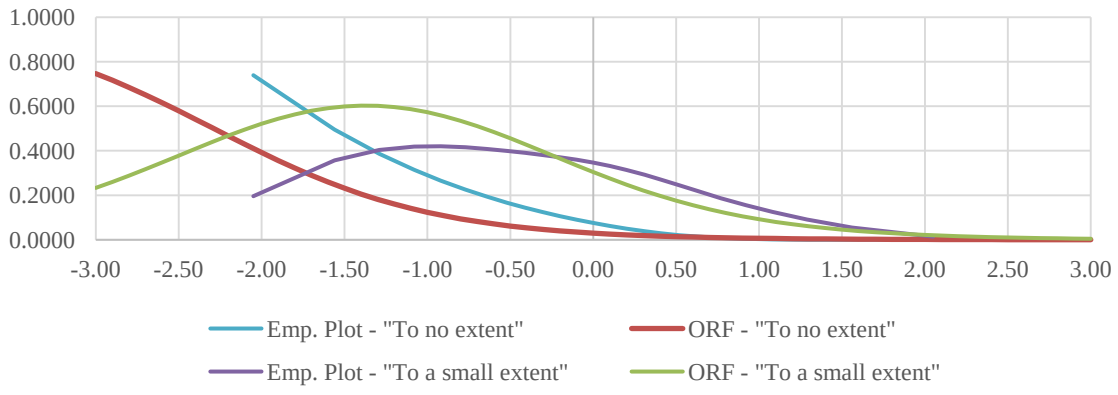

Figure 4.4. Empirical Plots and Option Response Functions (ORF's)

except for one item ("How much do you know about current renewable energy programs?") at the upper response options ("A fair amount" and "A lot"). Empirical fit curves indicated that members of the population that had the highest estimated levels of environmental sustainability engagement were more probable to respond "A fair amount" instead of "A lot". This could mean one of two things. First, the fourth response option ("A lot") might not be needed with this sample, but this is unlikely since this response option was chosen by 9% of the sample. This leaves the second option, which is that the item threshold between these two response options should be higher on the latent trait continuum for this sample. Besides these response options for these two items, empirical model-fit was strong, indicating invariance of the models and, thus, generalizability of the STEMSEI.

Finally, the last validity test pertained to statistically significant differences among the economic and environmental scale scores. Given that the IRT EAP estimated subscale scores for both the economic and environmental models had too few items, differences along the subscales (the emotional, cognitive (know), and cognitive (solve) scales) was not assessed. Since only the general score for each scale was being considered for differences among groups, ANOVA analyses were performed instead of MANOVA's. Five separate ANOVA's along race, gender, present religion, classification, and STEM major were performed on each of the general scale scores for the environmental and economic scales. Before the ANOVA analysis along STEM majors could be performed, three STEM major classifications had to be dropped from the analysis due to small subsamples ($n < 12$), including chemistry, computer science, and physics and astronomy. Five separate ANOVA analyses for the economic and environmental scales revealed statistically significant differences based on STEM major for both scales (as hypothesized) (see Table 4.16). Additionally, statistically significant differences based on present religion occurred in the environmental scale scores (as hypothesized) (see Table 4.16). As expected, further post-hoc tests revealed Christians scored statistically significantly lower ($M = -.034$, $SD = .864$) than non-Christians ($M = .233$, $SD = .805$) on the environmental general scale, $t(259) = -2.434$, $p = .016$, with Cohen's d = -.32.

Table 4.16

One-way ANOVA's for First Round Economic and Environmental General Scale Scores

	Variable	*F*	*df*	*p*	*B. H. p**	η^2
Economic General Scale	Race	.352	1, 317	.553	.553	.091
	Gender	4.386	1, 331	.037	.062	.013
	Present Religion	5.071	1, 260	.025	.062	.019
	Classification	1.624	1, 330	.203	.254	.005
	STEM major	**6.829**	**7, 284**	**< .001**	**.005**	**.147**
Environmental General Scale	Race	0.005	1, 317	.946	.946	.000
	Gender	0.015	1, 331	.902	.946	.000
	Present Religion	**5.925**	**1, 260**	**.016**	**.040**	**.022**
	Classification	0.007	1, 330	.934	.946	.000
	STEM major	**8.106**	**7, 284**	**< .001**	**.005**	**.170**

Note. * = Benjamini-Hochberg correction values.

Phase 3, step 9, round 1 results: Qualitative piloting in the sample. In this portion of phase 3, the semi-structured interview protocol (see Appendix F) was used to gather information from five post-secondary students on why they might perceive the economic, environmental, and social domains of sustainability as distinct and partially separate.

All five students commented that sustainability was also about making resource use more efficient. Four out of five of the post-secondary students interviewed also said they saw sustainability as an environmental issue. Only one student (a senior engineering major) explained sustainability as a multifaceted issue concerning economic, environmental, and social domains. This student was also able to identify and explain connections between these three domains, such as how social desires drive economic developments, which in turn drive which resources we extract from the environment. This cycle of product development was further explained by the student in nuanced ways, such as the finite amount of resources (economic, environmental, and social capitol) that we have and that must be appropriated in a way to keep the system in "equilibrium."

With respect to the specific ties the environment has to sustainability, many students echoed their previous ideas that sustainability was “exactly” the environment. The STEM education major explained that it was “hard to think of it as anything else.” The other students further explained that the environment was the source for all resources we use or consume. Beyond this, only the senior engineering major was able to explain more distinct connections between the environment and sustainability.

On the other hand, all students related various environmental concerns or limitations (i.e., finite resources such as fossil fuels) to economic concerns with respect to sustainability when asked how economics connected with sustainability. One common connection was how certain economic influences force or entice consumers to participate in the market in specific ways. For example, one engineering student highlighted that gas prices could rise higher because “they know we’ll pay it.” Another student commented how American culture influences people to desire unsustainable housing options in suburban and urban locations. The senior engineering major added that “money talks”, and that sustainable products will not be pursued economically until consumers are willing to pay more for them so that companies can meet or exceed current profit margins; this sentiment was also echoed by the economics major.

Connections between society and sustainability focused mostly on social needs/desires for certain products or services and how that influences what resources we harvest from the environment. In this context, two students saw society as the main problem for sustainability, while two others saw it as the solution for sustainability. The economics major expressed a very different view, commenting that the political aspect of society is currently a problem for sustainability because of the economic limitations that imposes on businesses. He further explained that environmental regulation of businesses by the government should be abolished, and that the market should ultimately decide in what ways sustainability is pursued in our society. In contrast, all of the other students, regardless whether they saw society as the solution or problem to sustainability, expressed that society has natural limitations on it from the environment and that those limitations naturally constrain aspects of society (e.g., population growth rates, consumer patterns, etc.). Of particular interest were one student's comments that she felt disconnected with some of her Christian peers due to her views on society and the

environment. She expressed the belief that some Christians disbelieved in climate change because of their faith. She departed from this belief and, instead, wanted to address environmental issues she believes are caused by humans (e.g., climate change) as an expression of her faith (i.e., to be "grateful for the Earth" because God made it for us).

When asked for differences in how economics, the environment, and society connected to sustainability, all participants agreed that there were differences, but the description of those differences varied across the sample. The mathematics major explained that the environment was “the stage” for sustainability while society was “the actor” and economics helped write “the script.” Somewhat similarly, the STEM education major expressed that she felt impacts for sustainability were observed first in the environment, then economically, and finally in society. The freshman engineering major explained that they were all “separate but related.”

Finally, there were varying responses amongst the participants in terms of their capacity to address the needs of sustainability in economic, environmental, and social contexts. Only the engineering majors felt prepared to address the environmental needs of sustainability, while only the economics major felt prepared to address the economic needs of sustainability. No participant felt their major prepared them to address the social needs of sustainability, though the mathematics major did express some possibilities using network theory to manage resources. Finally, none of the participants felt motivated to work in a position that addressed the needs of sustainability.

These results helped inform the item writing process in several ways. First, there was an inherent connection between the three domains of sustainability, which meant that item writing should continue along the delineation of the three domains of sustainability. Second, it established that differing engagement levels should be expected based on academic major and that disciplinary expertise should indicate higher levels of engagement in some cases (e.g., economics majors being more engaged in the economic domain of sustainability, while engineering majors should be more engaged in the environmental domain of sustainability); this needed to be further reflected in the item content to differentiate between various majors.

Phase 3, step 7, round 2 results: Preparing for instrument piloting and pilot administration. In the second round of piloting, 193 respondents participated from

PSIA, while 60 participated from PSIB. This resulted in a total of 253 responses for round 2. Since it was assumed that no person participated in both the first and second round of piloting, this brought cumulative response rates to 3.2% and 4.3% for the first and second institutions respectively. Only 1.5% of the overall data were missing in this piloting round. Missing data was again assumed to be missing not at random (MNAR). As before, all response categories were used to varying degrees across all items. Please refer to Table 3.5 for additional information on the sample gathered in round 2.

Phase 3, step 8a, round 2 results: Calculate reliability. Across all STEMSEI items in the complete sample, Cronbach's alpha was $\alpha = .928$. Across the economic, environmental, and social items, reliability was estimated to be, in order, $\alpha = .892$, $\alpha = .887$, and $\alpha = .866$. When compared to reliability estimates in the previous round (see Table 4.4), round 2 estimates of reliability were improvements for the economic and environmental items, but was a substantial improvement for the social items. Per Kline's (2000) recommendations, these reliability estimates were adequate to proceed. No items were dropped.

Phase 3, step 8b, round 2 results: Item analysis. In round 2 of this step of phase 3, the revised STEMSEI items were assessed for item functioning using the same methods as before. The following results first detail how item analysis transitioned from factor analysis results to item response theory results (i.e., quantitative piloting).

Round 2 factor analysis results. Since the factor structures of the economic and environmental scales were observed in the first piloting round, a confirmatory bifactor analysis approach was used for these scales. Results for the initial confirmatory bifactor analysis of the economic scale indicated strong fit on the CFI and RMSEA fit statistics, while initial confirmatory bifactor analysis results on the environmental scale were poorer (see Table 4.17). To improve model fit for the environmental scales, items with poor fit were dropped to produce a final set of items for the environmental scale, resulting in a final bifactor extraction for the environmental scale (see Table 4.17). Statistically significant results of the chi-square test of model fit were disregarded given the statistic's sensitivity in samples of larger size (Meyers, Gamst, & Guarino, 2006).

Unlike the economic and environmental scales, a factor structure was not identified for the social scale in the first round. However, using an exploratory factor

Table 4.17
Final Bifactor Extractions for the Economic and Environmental Scales

		Chi-Square Test			CFI	RMSEA
		χ^2	*df*	*p*		
Economic Scale	Bifactor Extraction	404.556	187	< .001	.967	.068
Environmental Scale	Initial Bifactor Extraction	952.956	273	< .001	.892	.099
	Final Bifactor Extraction	346.650	150	< .001	.959	.072

Note. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

analysis (EFA) approach on the social items in round 2, a factor structure mirroring that observed in the economic and environmental scales was identified (see Table 4.18). While the CFI statistic indicated model-fit, the χ^2 model fit statistic was statistically significant and the estimated RMSEA was greater than .08 (see Table 4.19), indicating lack of model-fit. However, as before with the economic and environmental scales, a corresponding bifactor model across the social items improved model-fit (see Table 4.21). As with the environmental scale in the previous round, the social scale experienced a large improvement in fit in the bifactor model as compared to the extracted EFA model, bringing fit statistics within acceptable ranges for the CFI and RMSEA. In short, the bifactor model seemed to indicate strong model-fit across all three scales in round 2.

Round 2 item response theory results. Using the multidimensional GPCM, multidimensional GRM, and the bifactor IRT model, initial models were calculated for all three scales. Again, rather than detail the item omission process for all models, evidence of the best model is presented first. A detailing of the item omission process for that model then follows.

Use of the AIC Difference statistic (see Burnham & Anderson, 2002) showed that the fit of the bifactor IRT models on the economic, environmental, and social scales were vast improvements over the multidimensional GRM and GPCM (see Table 4.20). Due to this, only the bifactor IRT model is further considered.

Table 4.18

Factor Patterns of an Exploratory Factor Analysis of the Social Items

	Item	Factor Patterns			Engagement Type
		1	2	3	
Social Scale	1	**0.819**	-0.045	-0.114	Emotional
	2	**0.511**	0.277	-0.01	Emotional
	3	**0.899**	-0.055	-0.002	Emotional
	4	**0.311**	0.302	-0.168	Emotional
	5	**0.505**	0.088	0.204	Emotional
	6	**0.417**	0.136	0.297	Emotional
	7	-0.065	**0.700**	-0.151	Cognitive (Know)
	8	0.014	**0.761**	-0.02	Cognitive (Know)
	9	-0.078	**0.429**	0.133	Cognitive (Know)
	10	0.123	**0.581**	0.148	Cognitive (Know)
	11	0.185	**0.711**	0.144	Cognitive (Know)
	12	-0.178	**0.615**	-0.087	Cognitive (Know)
	13	-0.026	**0.788**	0.027	Cognitive (Know)
	14	0.134	**0.624**	0.045	Cognitive (Know)
	15	0.043	0.075	**0.282**	Cognitive (Solution)
	16	-0.045	-0.045	**0.879**	Cognitive (Solution)
	17	-0.09	-0.079	**0.991**	Cognitive (Solution)
	18	-0.002	0.031	**0.863**	Cognitive (Solution)
	19	0.209	0.205	**0.646**	Cognitive (Solution)
	20	0.226	0.247	**0.598**	Cognitive (Solution)

Table 4.19

Corresponding Bifactor Model Extraction Along the Social Items

	Extraction	Chi-square Test			CFI	RMSEA
		χ^2	*df*	*p*		
Social Scale	EFA	561.295	150	< .001	.914	.104
	Bifactor	355.767	150	< .001	.950	.074

Note. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation.

Table 4.20
GPCM, GRM, and Bifactor IRT Models AIC Statistics and AIC Difference Statistics

	Model	AIC statistic	AIC Difference statistic
Economic Scale	3-factor GPCM	10645.43	-
	3-factor GRM	10612.44	32.99
	3-factor GPCM	10645.43	-
	Bifactor Model	10316.46	328.97
	3-factor GRM	10612.44	-
	Bifactor Model	10316.46	295.98
Environmental Scale	3-factor GPCM	9587.49	-
	3-factor GRM	9881.78	294.29
	3-factor GPCM	9587.49	-
	Bifactor Model	9568.63	18.86
	3-factor GRM	9881.78	-
	Bifactor Model	9568.63	313.15
Social Scale	3-factor GPCM	9905.16	-
	3-factor GRM	9887.56	17.6
	3-factor GPCM	9905.16	-
	Bifactor Model	9867.34	37.82
	3-factor GRM	9887.56	-
	Bifactor Model	9867.34	20.22

For each bifactor model, only one item had to be dropped per scale to attain strong item-level fit. The economic scale estimated one item ("How much concern do you have about the stability of economies in other countries?") with a low item slope parameter (a = 0.24) along the general factor. In a similar fashion, the bifactor IRT model for the social scale estimated a low item slope parameter (a = 0.37) for one item ("To what extent could your major measure various effects from the global population rate?") for the second cognitive engagement factor (solving). Finally, one item ("How much concern do you have about the rate at which humans consume non-renewable resources?") was removed from the environmental scale due to a large item slope parameter (a = 3.35) along the emotional factor. For each scale, the standardized LD χ^2

statistic was less than 10 for each item pair, as recommended by Cai, du Toit, and Thissen (2011, p. 77), after omitting the items above. The S-χ^2 item level diagnostic statistic also indicated item-level fit in the resultant models for each item (see Tables 4.21, 4.22, and 4.23). Final bifactor IRT models using round 2 data for the economic, environmental, and social scales can be found in Tables 4.21, 4.22, and 4.23, respectively.

When considering differential item functioning (DIF) analyses, given that the second round sample was 253, separating the sample into subsamples based on demographic variables presents lower statistical power for DIF analyses. Moreover, some response categories were not used across several subsamples, which interferes with two aspects of the DIF analyses. First, insufficient spread across all response options creates issues estimating item parameters with multidimensional graded response models (De Ayala, 1994). By conducting the DIF analyses and separating the sample into different subsamples, this manifested the insufficient spread in option responses. Second, the DIF analyses are calculated via Wald tests (see Lord, 1980), which are chi-squared tests. These are impossible to calculate with cell values of 0 (i.e., instances of a response option not being selected in a subsample). This occurred through a few of the DIF analyses. Due to these limiting factors, DIF analyses on the second round of data should be interpreted with caution.

In the DIF analyses of the economic scale, one item ("How much concern do you have about the United States economy being stable?") had to be dropped from all of the DIF analyses (i.e., for race, present religion, gender, and classification). This was because in each DIF analysis (i.e., for race, present religion, gender, and classification), one of the subsamples did not respond "no concern" to the item, leaving an empty cell in the chi-square test utilized in the Wald test (see Lord, 1980). Similarly, two items ("How much concern do you have about the United States economy being stable?" and "how much concern do you have about businesses being efficient in product manufacturing?") had to be dropped from both the race DIF analysis and the present religion DIF analysis as well. After dropping the corresponding items from each DIF analysis, no DIF was detected on the economic scale (see Table 4.24).

Table 4.21

Final Bifactor IRT Model Parameters for Economic Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistics			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p*	*B.H.p**
1	1.2 (0.19)	1.83 (0.23)	0 (-)	0 (-)	4.30 (0.42)	0.65 (0.21)	-3.04 (0.32)	66.65	57	0.179	.458
2	0.96 (0.18)	1.65 (0.23)	0 (-)	0 (-)	5.56 (0.62)	1.85 (0.23)	-1.28 (0.21)	58.10	53	0.293	.458
3	1.08 (0.18)	1.42 (0.20)	0 (-)	0 (-)	4.73 (0.48)	0.76 (0.19)	-2.36 (0.26)	59.20	51	0.201	.458
4	1.11 (0.19)	1.85 (0.24)	0 (-)	0 (-)	4.20 (0.41)	1.20 (0.22)	-1.90 (0.24)	64.35	60	0.326	.458
5	0.86 (0.18)	1.5 (0.24)	0 (-)	0 (-)	6.79 (1.07)	3.17 (0.32)	0.03 (0.18)	49.42	44	0.265	.458
6	0.72 (0.17)	1.6 (0.23)	0 (-)	0 (-)	4.92 (0.52)	2.00 (0.24)	-1.18 (0.20)	61.02	57	0.333	.458
7	1.57 (0.20)	0.92 (0.17)	0 (-)	0 (-)	4.02 (0.37)	1.81 (0.23)	-1.53 (0.22)	65.76	49	0.055	.385
8	2.96 (0.31)	0 (-)	1.19 (0.22)	0 (-)	1.78 (0.35)	-2.87 (0.37)	-7.78 (0.85)	47.10	38	0.148	.458
9	2.58 (0.29)	0 (-)	1.13 (0.21)	0 (-)	2.28 (0.34)	-2.65 (0.36)	-6.00 (0.62)	49.61	42	0.195	.458
10	2.3 (0.25)	0 (-)	1.26 (0.20)	0 (-)	2.46 (0.32)	-2.18 (0.31)	-6.03 (0.61)	68.92	39	**0.002 ****	**.042****
11	1.77 (0.21)	0 (-)	1.43 (0.20)	0 (-)	3.40 (0.36)	-0.59 (0.23)	-4.38 (0.41)	51.67	48	0.332	.458
12	1.06 (0.25)	0 (-)	3.02 (0.32)	0 (-)	2.91 (0.37)	-1.88 (0.35)	-7.01 (0.79)	63.00	56	0.242	.458
13	1.01 (0.24)	0 (-)	2.97 (0.33)	0 (-)	2.94 (0.37)	-1.73 (0.35)	-6.70 (0.76)	65.88	54	0.129	.458
14	0.6 (0.19)	0 (-)	2.04 (0.25)	0 (-)	0.91 (0.22)	-2.64 (0.32)	-6.56 (0.76)	75.89	53	0.021	.221
15	1.91 (0.24)	0 (-)	0 (-)	0.66 (0.17)	2.50 (0.27)	0.41 (0.22)	-1.71 (0.26)	50.76	59	0.769	.769
16	2.63 (0.30)	0 (-)	0 (-)	1.29 (0.22)	3.63 (0.39)	0.55 (0.27)	-2.27 (0.33)	49.60	49	0.450	.525
17	1.43 (0.20)	0 (-)	0 (-)	0.91 (0.17)	1.03 (0.20)	-0.42 (0.20)	-1.78 (0.23)	60.57	63	0.564	.623
18	0.92 (0.26)	0 (-)	0 (-)	2.87 (0.33)	1.86 (0.34)	-1.26 (0.29)	-3.84 (0.43)	77.12	67	0.186	.458

Table 4.21 (cont.)
Final Bifactor IRT Model Parameters for Economic Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistics			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p*	*B.H.p**
19	2.16 (0.26)	0 (-)	0 (-)	1.13 (0.19)	3.41 (0.35)	0.08 (0.23)	-2.57 (0.31)	55.85	53	0.368	.458
20	0.72 (0.24)	0 (-)	0 (-)	2.77 (0.35)	1.65 (0.33)	-1.27 (0.27)	-3.43 (0.39)	75.32	72	0.371	.458
21	1.99 (0.30)	0 (-)	0 (-)	2.05 (0.28)	1.87 (0.27)	-0.85 (0.29)	-3.36 (0.44)	54.56	58	0.605	.635

Note. Values in ( ) are parameter estimate standard errors. a = item slope; b = item location. These values are for a logistic scale.
* Benjamini-Hochberg correction values.
** empirical fit analysis revealed strong fit on this item; considered false positive.

Table 4.22
Final Bifactor IRT Model Parameters for Environmental Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistics			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p*	*B.H.p**
1	1.58 (0.28)	2.39 (0.3)	0 (-)	0 (-)	8.62 (1.22)	4.89 (0.52)	0.82 (0.27)	49.54	33	0.032	.213
2	1.23 (0.25)	1.88 (0.25)	0 (-)	0 (-)	6.4 (0.74)	3.91 (0.4)	0.48 (0.23)	44.59	33	0.086	.356
3	1.03 (0.22)	1.86 (0.24)	0 (-)	0 (-)	4.12 (0.41)	1.85 (0.26)	-0.31 (0.21)	70.69	56	0.089	.356
4	1.81 (0.31)	1.66 (0.24)	0 (-)	0 (-)	5.34 (0.55)	2.74 (0.33)	-0.67 (0.23)	47.25	43	0.302	.549
5	1.36 (0.26)	2.35 (0.31)	0 (-)	0 (-)	5.59 (0.62)	2.4 (0.35)	-1.67 (0.27)	58.09	48	0.151	.378
6	0.58 (0.18)	1.3 (0.19)	0 (-)	0 (-)	2.1 (0.22)	-0.28 (0.17)	-2.75 (0.27)	**90.20**	**57**	**0.003****	**.030****
7	1.96 (0.35)	0 (-)	1.54 (0.33)	0 (-)	4.37 (0.44)	0.52 (0.23)	-3.14 (0.34)	37.35	45	0.785	.823
8	1.93 (0.36)	0 (-)	1.34 (0.32)	0 (-)	4.59 (0.47)	0.58 (0.23)	-3.04 (0.34)	51.86	44	0.194	.431

Table 4.24 (cont.)
Final Bifactor IRT Model Parameters for Environmental Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistics			
	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p**	*B.H.p**
9	1.46 (0.24)	0 (-)	0.74 (0.23)	0 (-)	2.94 (0.29)	0.12 (0.18)	-2.08 (0.24)	68.76	57	0.137	.378
10	1.47 (0.31)	0 (-)	2.14 (0.3)	0 (-)	3.08 (0.38)	-0.1 (0.24)	-2.87 (0.35)	56.46	61	0.641	.823
11	2.07 (0.29)	0 (-)	2.59 (0.42)	0 (-)	6.34 (0.66)	1.82 (0.31)	-2.16 (0.36)	42.68	49	0.726	.823
12	1.08 (0.21)	0 (-)	1.38 (0.24)	0 (-)	5.46 (0.6)	1.98 (0.24)	-1.33 (0.21)	53.02	47	0.253	.506
13	1.31 (0.29)	0 (-)	1.69 (0.3)	0 (-)	4.35 (0.41)	1.4 (0.23)	-1.37 (0.24)	51.72	50	0.408	.583
14	1.71 (0.26)	0 (-)	0 (-)	2.13 (0.26)	3.12 (0.35)	1.25 (0.26)	-1.24 (0.26)	65.74	54	0.131	.378
15	1.91 (0.29)	0 (-)	0 (-)	2.16 (0.26)	4.18 (0.44)	1.6 (0.28)	-0.83 (0.26)	38.49	47	0.808	.823
16	1.67 (0.29)	0 (-)	0 (-)	1.9 (0.25)	2.93 (0.33)	1 (0.25)	-1.19 (0.25)	45.47	54	0.790	.823
17	1.64 (0.29)	0 (-)	0 (-)	2.65 (0.33)	2.79 (0.37)	0.9 (0.28)	-1.54 (0.3)	46.17	56	0.823	.823
18	1.65 (0.29)	0 (-)	0 (-)	1.33 (0.21)	2.36 (0.28)	0.85 (0.22)	-1.1 (0.22)	56.05	53	0.361	.580
19	1.05 (0.24)	0 (-)	0 (-)	2.14 (0.28)	2.2 (0.3)	0.38 (0.23)	-1.75 (0.26)	67.94	65	0.377	.580

Note. Values in ( ) are parameter estimate standard errors. a = item slope; b = item location. These values are for a logistic scale.
* Benjamini-Hochberg correction values.
** empirical fit analysis revealed strong fit on this item; considered false positive.

Table 4.23

Final Bifactor IRT Model Parameters for Social Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistic			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p**	*B.H.p**
1	1.05 (0.25)	2.51 (0.33)	0 (-)	0 (-)	5.23 (0.58)	2.76 (0.37)	-0.73 (0.26)	51.03	47	0.318	.481
2	1.33 (0.22)	0.97 (0.2)	0 (-)	0 (-)	5.2 (0.57)	2.74 (0.28)	-0.21 (0.18)	55.07	39	0.045	.269
3	1.59 (0.28)	2.49 (0.33)	0 (-)	0 (-)	5.54 (0.61)	2.04 (0.33)	-1.15 (0.28)	61.94	48	0.085	.269
4	0.75 (0.17)	0.49 (0.15)	0 (-)	0 (-)	3.12 (0.29)	0.82 (0.15)	-0.93 (0.16)	53.45	56	0.573	.640
5	1.35 (0.23)	0.65 (0.19)	0 (-)	0 (-)	5.26 (0.61)	3.15 (0.31)	-0.96 (0.2)	34.39	33	0.403	.513
6	1.83 (0.27)	0.44 (0.19)	0 (-)	0 (-)	4.82 (0.51)	2.25 (0.28)	-0.77 (0.19)	50.24	40	0.128	.314
7	0.87 (0.18)	0 (-)	0.98 (0.19)	0 (-)	3.73 (0.36)	0.65 (0.17)	-2.02 (0.21)	37.47	47	0.839	.839
8	1.29 (0.23)	0 (-)	2.05 (0.3)	0 (-)	1.4 (0.25)	-2.5 (0.31)	-5.99 (0.68)	54.64	41	0.075	.269
9	0.68 (0.16)	0 (-)	0.87 (0.18)	0 (-)	1.87 (0.2)	-0.87 (0.16)	-3.45 (0.32)	44.83	51	0.716	.756
10	1.82 (0.25)	0 (-)	0.81 (0.2)	0 (-)	3 (0.31)	-0.49 (0.2)	-3.31 (0.33)	62.87	48	0.073	.269
11	2.32 (0.29)	0 (-)	1.7 (0.27)	0 (-)	2 (0.3)	-2.26 (0.3)	-5.31 (0.53)	38.87	38	0.432	.513
12	0.66 (0.16)	0 (-)	0.94 (0.18)	0 (-)	2.01 (0.2)	-0.42 (0.16)	-2.84 (0.27)	61.89	52	0.164	.346
13	1.5 (0.24)	0 (-)	2.05 (0.29)	0 (-)	1.19 (0.25)	-3.07 (0.34)	-5.7 (0.6)	43.61	39	0.281	.481
14	1.27 (0.22)	0 (-)	1.65 (0.26)	0 (-)	0.04 (0.2)	-2.68 (0.28)	-4.63 (0.46)	58.59	42	0.046	.269
15	0.62 (0.16)	0 (-)	0 (-)	0.55 (0.14)	1.39 (0.17)	0.33 (0.15)	-1.23 (0.17)	67.39	63	0.329	.481
16	1.51 (0.27)	0 (-)	0 (-)	2.69 (0.3)	0.67 (0.29)	-2.21 (0.34)	-5.35 (0.55)	53.35	52	0.423	.513
17	1.93 (0.29)	0 (-)	0 (-)	3.11 (0.33)	1.73 (0.36)	-1.59 (0.35)	-5.25 (0.55)	68.68	53	0.072	.269
18	1.86 (0.26)	0 (-)	0 (-)	2.4 (0.29)	1.55 (0.32)	-1.66 (0.29)	-4.7 (0.46)	62.36	51	0.132	.314

Table 4.23 (cont.)
Final Bifactor IRT Model Parameters for Social Scale

	Item Slope Parameters By Factor/Subfactor				Item Intercept			S-χ^2 Item Level Statistic			
Item	General (a)	Emotional (a_1)	Cognitive (Knowing) (a_2)	Cognitive (Solving) (a_3)	c_1	c_2	c_3	χ^2	*df*	*p**	*B.H.p**
19	1.65 (0.23)	0 (-)	0 (-)	0.67 (0.17)	1.27 (0.2)	-0.98 (0.21)	-2.78 (0.29)	61.33	53	0.202	.384

Note. Values in ( ) are parameter estimate standard errors. a = item slope; b = item location. These values are for a logistic scale.
* assessed at the α = .05/19 = .003 level.

In the environmental scale DIF analyses, the item "how much concern do you have about pollution of the environment?" also had to be dropped from all DIF analyses due to empty cells, specifically no responses to the "no concern" option. In the present religion and gender DIF analyses, two additional items ("How much concern do you have about finding renewable resources?" and "I find it interesting to know what my own carbon footprint is.") were dropped for similar reasons. In the gender DIF, the item "I would be comfortable having a limited amount of water I could use daily" exhibited non-uniform DIF. A secondary DIF analysis with all other items anchored still resulted in DIF on this item, $\chi^2(4, N = 118) = 17.0$, $p = .0019$, which was again indicated to be non-uniform DIF, $\chi^2(1, N = 118) = 9.5$, $p = .0020$. Of particular interest in this item is that a near-zero, negative item slope parameter was estimated for males in the DIF analysis. Since this item indicated model-fit in the bifactor IRT model, it was not omitted from further analyses but may require further scrutiny in future use of the STEMSEI. No other positive DIF results were observed in the environmental scale (see Table 4.23).

In the social scale DIF analyses, several items were dropped from the present religion and gender DIF analyses due again to empty response cells. Ultimately however, the DIF analyses of the social scale indicated no statistically significant DIF across race, present religion, gender, or classification (see Table 4.24).

Both marginal reliability and the standard error of approximation (SEE) greatly improved from the previous round. The final bifactor IRT models for each scale (economic, environmental, and social) produced scores accurate to about from ±.38 to

Table 4.24

Phase 3, Round 2: Differential Item Functioning Tests for Economic Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Race	1	10.5	4	0.033	.627	2.7	1	0.099	-	7.8	3	0.051	-
	2	-	-	-	-	-	-	-	-	-	-	-	-
	3	4.1	4	0.388	.853	0.2	1	0.684	-	4	3	0.265	-
	4	3.9	4	0.427	.853	1.4	1	0.242	-	2.5	3	0.479	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	0.8	4	0.938	.938	0	1	0.896	-	0.8	3	0.853	-
	7	5.4	4	0.253	.801	1.6	1	0.204	-	3.8	3	0.291	-
	8	2.4	4	0.669	.938	0.7	1	0.388	-	1.6	3	0.655	-
	9	3.7	4	0.449	.853	0.5	1	0.477	-	3.2	3	0.363	-
	10	2.1	4	0.720	.938	0.2	1	0.635	-	1.9	3	0.602	-
	11	5.7	4	0.225	.801	0.7	1	0.394	-	5	3	0.176	-
	12	0.8	4	0.932	.938	0	1	0.851	-	0.8	3	0.847	-
	13	4	4	0.402	.853	0.4	1	0.527	-	3.6	3	0.305	-
	14	1.2	4	0.873	.938	0	1	0.989	-	1.2	3	0.745	-
	15	1.2	4	0.873	.938	0	1	0.903	-	1.2	3	0.750	-
	16	1.8	4	0.767	.938	0.8	1	0.382	-	1.1	3	0.785	-
	17	2	4	0.733	.938	0.2	1	0.641	-	1.8	3	0.616	-
	18	5.8	4	0.214	.801	0.9	1	0.341	-	4.9	3	0.179	-
	19	2.3	4	0.674	.938	0.1	1	0.790	-	2.3	3	0.519	-
	20	6.4	4	0.171	.801	0.1	1	0.800	-	6.3	3	0.097	-
	21	8.6	4	0.073	.684	2	1	0.163	-	6.6	3	0.085	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.24 (cont.)

Phase 3, Round 2: Differential Item Functioning Tests for Economic Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Present Religion	1	3.8	4	0.439	.999	0.5	1	0.459	-	3.2	3	0.360	-
	2	-	-	-	-	-	-	-	-	-	-	-	-
	3	1.3	4	0.856	.999	0.1	1	0.703	-	1.2	3	0.756	-
	4	2.7	4	0.606	.999	0	1	0.973	-	2.7	3	0.437	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	1.9	4	0.749	.999	0	1	0.858	-	1.9	3	0.595	-
	7	2	4	0.736	.999	0.2	1	0.654	-	1.8	3	0.616	-
	8	0.1	4	0.999	.999	0	1	0.855	-	0.1	3	0.995	-
	9	0.7	4	0.948	.999	0	1	0.946	-	0.7	3	0.868	-
	10	1.1	4	0.892	.999	0.7	1	0.406	-	0.4	3	0.935	-
	11	0.5	4	0.969	.999	0	1	0.919	-	0.5	3	0.911	-
	12	2.7	4	0.619	.999	0.4	1	0.529	-	2.3	3	0.522	-
	13	1.4	4	0.841	.999	0.6	1	0.424	-	0.8	3	0.854	-
	14	3.4	3	0.333	.999	0.3	1	0.570	-	3.1	2	0.215	-
	15	8.1	4	0.087	.999	0.3	1	0.612	-	7.9	3	0.049	-
	16	2.5	4	0.640	.999	0	1	0.896	-	2.5	3	0.474	-
	17	5.1	4	0.280	.999	1.2	1	0.278	-	3.9	3	0.273	-
	18	0.4	4	0.986	.999	0	1	0.893	-	0.3	3	0.954	-
	19	2.3	4	0.678	.999	1.2	1	0.284	-	1.2	3	0.761	-
	20	1.3	4	0.859	.999	0	1	0.935	-	1.3	3	0.728	-
	21	1.6	4	0.801	.999	0.3	1	0.587	-	1.3	3	0.719	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.24 (cont.)

Phase 3, Round 2: Differential Item Functioning Tests for Economic Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Gender	1	5.9	4	0.209	.888	2	1	0.154	-	3.9	3	0.279	-
	2	0.7	4	0.949	.996	0.5	1	0.484	-	0.2	3	0.973	-
	3	1.6	4	0.806	.996	0.1	1	0.744	-	1.5	3	0.681	-
	4	1	4	0.907	.996	0.3	1	0.570	-	0.7	3	0.874	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	3.8	4	0.440	.978	1.6	1	0.209	-	2.2	3	0.538	-
	7	4.2	4	0.382	.978	0.1	1	0.746	-	4.1	3	0.254	-
	8	1.6	3	0.659	.996	0	1	0.977	-	1.6	2	0.449	-
	9	1.1	4	0.898	.996	0.1	1	0.819	-	1	3	0.796	-
	10	1.4	4	0.837	.996	0	1	0.951	-	1.4	3	0.697	-
	11	3.8	4	0.440	.978	0.6	1	0.423	-	3.1	3	0.375	-
	12	4.4	3	0.222	.888	0.2	1	0.678	-	4.2	2	0.120	-
	13	10.9	4	0.027	.540	0.4	1	0.547	-	10.6	3	0.014	-
	14	6.1	3	0.108	.888	2.2	1	0.144	-	3.9	2	0.142	-
	15	2.7	4	0.618	.996	0	1	0.952	-	2.7	3	0.449	-
	16	2	4	0.735	.996	0	1	0.840	-	2	3	0.581	-
	17	0.2	4	0.996	.996	0	1	0.967	-	0.2	3	0.979	-
	18	1.4	4	0.853	.996	0	1	0.917	-	1.3	3	0.720	-
	19	3.8	4	0.438	.978	1.2	1	0.279	-	2.6	3	0.458	-
	20	5.8	4	0.219	.888	1.4	1	0.235	-	4.4	3	0.227	-
	21	3.1	4	0.547	.996	1.1	1	0.287	-	1.9	3	0.588	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.24 (cont.)

Phase 3, Round 2: Differential Item Functioning Tests for Economic Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Classification	1	0.7	4	0.949	.957	0.1	1	0.713	-	0.6	3	0.899	-
	2	5.3	4	0.263	.957	1.8	1	0.179	-	3.5	3	0.328	-
	3	3.9	4	0.419	.957	0	1	0.941	-	3.9	3	0.272	-
	4	2.4	4	0.656	.957	0.6	1	0.455	-	1.9	3	0.599	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	7.8	4	0.099	.957	0.1	1	0.709	-	7.7	3	0.053	-
	7	1.7	4	0.784	.957	0.2	1	0.642	-	1.5	3	0.678	-
	8	1.6	4	0.809	.957	0	1	0.900	-	1.6	3	0.663	-
	9	1	4	0.903	.957	0	1	0.870	-	1	3	0.797	-
	10	2.1	4	0.725	.957	0.3	1	0.617	-	1.8	3	0.613	-
	11	0.7	4	0.957	.957	0.1	1	0.785	-	0.6	3	0.902	-
	12	2.4	4	0.659	.957	1.9	1	0.175	-	0.6	3	0.903	-
	13	2.5	4	0.642	.957	1.1	1	0.299	-	1.4	3	0.698	-
	14	4.7	4	0.316	.957	0.2	1	0.665	-	4.6	3	0.208	-
	15	3.1	4	0.542	.957	1.2	1	0.272	-	1.9	3	0.595	-
	16	3.1	4	0.535	.957	0.1	1	0.796	-	3.1	3	0.381	-
	17	6.8	4	0.145	.957	0.4	1	0.552	-	6.5	3	0.091	-
	18	0.8	4	0.939	.957	0.6	1	0.431	-	0.2	3	0.982	-
	19	3.7	4	0.447	.957	0	1	0.930	-	3.7	3	0.296	-
	20	1	4	0.916	.957	0.6	1	0.443	-	0.4	3	0.947	-
	21	8.3	4	0.082	.957	4.6	1	0.031	-	3.6	3	0.306	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.25

Phase 3, Round 2: Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Race	1	-	-	-	-	-	-	-	-	-	-	-	-
	2	12	4	0.018	.324	4.9	1	0.027	-	7.1	3	0.070	-
	3	5.9	4	0.212	.984	0	1	0.892	-	5.8	3	0.120	-
	4	5.3	4	0.260	.984	3.1	1	0.079	-	2.2	3	0.532	-
	5	3.9	4	0.415	.984	1.2	1	0.266	-	2.7	3	0.441	-
	6	2.6	4	0.635	.984	0.1	1	0.752	-	2.5	3	0.484	-
	7	2.6	4	0.634	.984	0	1	0.857	-	2.5	3	0.471	-
	8	7.4	4	0.114	.984	0.4	1	0.554	-	7.1	3	0.069	-
	9	0.7	4	0.945	.984	0.1	1	0.738	-	0.6	3	0.888	-
	10	1.9	4	0.747	.984	0.7	1	0.390	-	1.2	3	0.753	-
	11	1.3	4	0.861	.984	0	1	0.956	-	1.3	3	0.730	-
	12	3.8	4	0.428	.984	2.3	1	0.127	-	1.5	3	0.682	-
	13	0.8	4	0.943	.984	0.3	1	0.572	-	0.4	3	0.931	-
	14	0.9	4	0.918	.984	0	1	0.990	-	0.9	3	0.815	-
	15	3.5	4	0.477	.984	1.3	1	0.263	-	2.3	3	0.522	-
	16	0.4	4	0.984	.984	0	1	0.915	-	0.4	3	0.946	-
	17	3.1	4	0.539	.984	0	1	0.825	-	3.1	3	0.382	-
	18	1.5	4	0.836	.984	0	1	0.839	-	1.4	3	0.704	-
	19	4.5	4	0.340	.984	0	1	0.898	-	4.5	3	0.212	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.25 (cont).

Phase 3, Round 2: Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Present Religion	1	-	-	-	-	-	-	-	-	-	-	-	-
	2	-	-	-	-	-	-	-	-	-	-	-	-
	3	9.7	4	0.047	.376	0.5	1	0.484	-	9.2	3	0.027	-
	4	5.3	4	0.262	.778	0.3	1	0.565	-	4.9	3	0.178	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	13.7	4	0.008	.128	0.5	1	0.495	-	13.2	3	0.004	-
	7	7.9	4	0.095	.507	2.2	1	0.137	-	5.7	3	0.128	-
	8	4.1	4	0.388	.778	0.2	1	0.647	-	3.9	3	0.270	-
	9	4.7	4	0.318	.778	1.9	1	0.173	-	2.9	3	0.415	-
	10	0.4	4	0.986	.986	0	1	0.893	-	0.3	3	0.952	-
	11	1.8	4	0.769	.986	0.9	1	0.345	-	0.9	3	0.820	-
	12	0.9	4	0.919	.986	0.7	1	0.400	-	0.2	3	0.973	-
	13	4.1	4	0.389	.778	0	1	0.856	-	4.1	3	0.252	-
	14	0.5	4	0.976	.986	0.2	1	0.674	-	0.3	3	0.962	-
	15	3.5	4	0.475	.844	2.1	1	0.144	-	1.4	3	0.710	-
	16	1.5	4	0.827	.986	0.7	1	0.397	-	0.8	3	0.854	-
	17	4.3	4	0.372	.778	0.1	1	0.799	-	4.2	3	0.242	-
	18	0.6	4	0.964	.986	0.1	1	0.815	-	0.5	3	0.911	-
	19	3	4	0.557	.891	0	1	0.904	-	3	3	0.393	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.25 (cont).

Phase 3, Round 2: Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Gender	1	-	-	-	-	-	-	-	-	-	-	-	-
	2	-	-	-	-	-	-	-	-	-	-	-	-
	3	4.6	4	0.330	.618	2.1	1	0.152	-	2.6	3	0.467	-
	4	6.9	4	0.144	.461	0.8	1	0.361	-	6	3	0.111	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	**16.9**	**4**	**0.002**	**.032**	**9.9**	**1**	**0.002**	-	7	3	0.072	-
	7	5.2	4	0.270	.617	0.1	1	0.738	-	5.1	3	0.166	-
	8	3.6	4	0.464	.617	0.7	1	0.392	-	2.9	3	0.414	-
	9	1.7	4	0.783	.895	0.1	1	0.715	-	1.6	3	0.658	-
	10	4.1	4	0.397	.617	0.6	1	0.429	-	3.4	3	0.329	-
	11	7.6	4	0.109	.436	0.5	1	0.476	-	7.1	3	0.070	-
	12	0.5	4	0.977	.977	0	1	0.909	-	0.5	3	0.929	-
	13	10.3	4	0.036	.192	0.2	1	0.677	-	10.1	3	0.018	-
	14	3.3	4	0.502	.618	1.4	1	0.239	-	2	3	0.582	-
	15	0.7	4	0.950	.977	0	1	0.998	-	0.7	3	0.870	-
	16	11.7	4	0.020	.160	0.2	1	0.660	-	11.5	3	0.009	-
	17	3.7	4	0.451	.618	0	1	0.983	-	3.7	3	0.298	-
	18	4.4	4	0.357	.618	1.5	1	0.226	-	2.9	3	0.405	-
	19	6	4	0.204	.544	0.2	1	0.666	-	5.8	3	0.123	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.25 (cont).

Phase 3, Round 2: Differential Item Functioning Tests for Environmental Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Classification	1	-	-	-	-	-	-	-	-	-	-	-	-
	2	4.8	4	0.313	.471	2	1	0.154	-	2.7	3	0.436	-
	3	8	4	0.090	.360	3.9	1	0.049	-	4.2	3	0.247	-
	4	8.9	4	0.063	.360	1	1	0.323	-	8	3	0.047	-
	5	7	4	0.136	.405	3.2	1	0.074	-	3.8	3	0.285	-
	6	1.8	4	0.781	.874	0.3	1	0.604	-	1.5	3	0.686	-
	7	1.2	4	0.874	.874	0.8	1	0.365	-	0.4	3	0.940	-
	8	4.8	4	0.314	.471	0.3	1	0.578	-	4.5	3	0.218	-
	9	16.1	4	0.003	.054	3.5	1	0.061	-	12.6	3	0.006	-
	10	1.4	4	0.841	.874	0.1	1	0.714	-	1.3	3	0.733	-
	11	4.8	4	0.313	.471	2	1	0.154	-	2.7	3	0.436	-
	12	11.6	4	0.021	.189	1.7	1	0.192	-	9.9	3	0.019	-
	13	4.8	4	0.305	.471	0.2	1	0.622	-	4.6	3	0.205	-
	14	6.3	4	0.176	.405	0.1	1	0.795	-	6.3	3	0.099	-
	15	1.4	4	0.848	.874	0	1	0.829	-	1.3	3	0.722	-
	16	3.5	4	0.485	.672	0.1	1	0.719	-	3.3	3	0.345	-
	17	6.3	4	0.180	.405	0	1	0.923	-	6.3	3	0.100	-
	18	7.8	4	0.100	.360	2.3	1	0.127	-	5.4	3	0.143	-
	19	1.6	4	0.807	.874	0	1	0.985	-	1.6	3	0.657	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.26

Phase 3, Round 2: Differential Item Functioning Tests for Social Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Race	1	5.2	4	0.266	.389	4.2	1	0.040	-	1	3	0.801	-
	2	2.3	4	0.686	.686	0.4	1	0.541	-	1.9	3	0.593	-
	3	5.8	4	0.212	.389	0.4	1	0.516	-	5.4	3	0.143	-
	4	3	4	0.560	.626	0.1	1	0.749	-	2.9	3	0.410	-
	5	10	4	0.041	.285	5.1	1	0.024	-	4.9	3	0.180	-
	6	2.4	4	0.657	.686	0	1	0.909	-	2.4	3	0.490	-
	7	6.5	4	0.165	.348	0.1	1	0.775	-	6.4	3	0.093	-
	8	8.6	4	0.071	.285	0.9	1	0.341	-	7.7	3	0.052	-
	9	8	4	0.091	.288	0.8	1	0.375	-	7.2	3	0.065	-
	10	7	4	0.136	.335	0	1	0.912	-	7	3	0.072	-
	11	3.6	4	0.459	.545	0.8	1	0.373	-	2.8	3	0.418	-
	12	5.7	4	0.225	.389	2.3	1	0.132	-	3.4	3	0.335	-
	13	5.3	4	0.260	.389	0.6	1	0.424	-	4.7	3	0.200	-
	14	10.4	4	0.035	.285	0	1	0.866	-	10.3	3	0.016	-
	15	8.5	4	0.075	.285	1	1	0.319	-	7.5	3	0.057	-
	16	3.7	4	0.453	.545	0.7	1	0.419	-	3	3	0.390	-
	17	8.6	4	0.072	.285	1.7	1	0.198	-	6.9	3	0.074	-
	18	6.9	4	0.141	.335	0.5	1	0.474	-	6.4	3	0.094	-
	19	3.6	4	0.459	.545	0.8	1	0.373	-	2.8	3	0.418	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.26 (cont.)

Phase 3, Round 2: Differential Item Functioning Tests for Social Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Present Religion	1	2.3	4	0.677	.988	0.1	1	0.819	-	2.3	3	0.519	-
	2	2.4	4	0.663	.988	0.3	1	0.572	-	2.1	3	0.556	-
	3	7.1	4	0.129	.988	0.6	1	0.446	-	6.6	3	0.087	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	-	-	-	-	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	-	-	-	-	-	-
	8	2.7	4	0.605	.988	0	1	0.911	-	2.7	3	0.438	-
	9	6.5	4	0.162	.988	2.9	1	0.088	-	3.6	3	0.308	-
	10	5.1	4	0.283	.988	0.9	1	0.356	-	4.2	3	0.241	-
	11	5.5	4	0.245	.988	0.1	1	0.755	-	5.4	3	0.147	-
	12	1.8	4	0.778	.988	1.1	1	0.304	-	0.7	3	0.870	-
	13	0.6	4	0.962	.988	0.1	1	0.726	-	0.5	3	0.922	-
	14	0.3	4	0.988	.988	0.2	1	0.695	-	0.2	3	0.982	-
	15	1.3	4	0.857	.988	0.1	1	0.733	-	1.2	3	0.750	-
	16	2.4	4	0.666	.988	0	1	0.834	-	2.3	3	0.506	-
	17	1.5	4	0.825	.988	0	1	0.970	-	1.5	3	0.681	-
	18	3.2	4	0.534	.988	0	1	0.984	-	3.2	3	0.370	-
	19	1.4	4	0.847	.988	0.2	1	0.647	-	1.2	3	0.759	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.26 (cont.)

Phase 3, Round 2: Differential Item Functioning Tests for Social Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Gender	1	8.8	4	0.065	.163	2.1	1	0.147	-	6.7	3	0.081	-
	2	-	-	-	-	-	-	-	-	-	-	-	-
	3	14.8	4	0.005	.056	0.2	1	0.622	-	14.5	3	0.002	-
	4	4.4	4	0.350	.438	3.5	1	0.060	-	0.9	3	0.822	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	6	-	-	-	-	-	-	-	-	-	-	-	-
	7	7.4	4	0.114	.193	4	1	0.046	-	3.5	3	0.326	-
	8	7.4	4	0.116	.193	4.6	1	0.032	-	2.8	3	0.429	-
	9	5.3	4	0.260	.355	0	1	0.912	-	5.3	3	0.152	-
	10	12.6	4	0.014	.563	10.5	1	0.001	-	2	3	0.569	-
	11	5.5	4	0.239	.355	5	1	0.026	-	0.6	3	0.908	-
	12	13	4	0.011	.563	1	1	0.326	-	12	3	0.007	-
	13	-	-	-	-	-	-	-	-	-	-	-	-
	14	2.6	4	0.630	.630	1.9	1	0.165	-	0.6	3	0.887	-
	15	3.8	4	0.429	.495	2.9	1	0.087	-	0.9	3	0.828	-
	16	9.8	4	0.043	.129	9.2	1	0.002	-	0.6	3	0.895	-
	17	12.3	4	0.015	.056	10.4	1	0.001	-	1.9	3	0.595	-
	18	8.4	4	0.079	.169	8	1	0.005	-	0.3	3	0.955	-
	19	2.6	4	0.621	.630	0	1	0.841	-	2.6	3	0.459	-

Note. B. H. p = Benjamini-Hochberg correction value.

Table 4.26 (cont.)
Phase 3, Round 2: Differential Item Functioning Tests for Social Items

		Omnibus Statistic				*Nonuniform DIF Statistic*				*Uniform DIF Statistic*			
	Item	Total X^2	*df*	*p*	*B.H. p*	X^2_a	*df*	*p*	*B.H. p*	$X^2_{c\|a}$	*df*	*p*	*B.H. p*
Classification	1	5.6	4	0.236	.403	0	1	0.943	-	5.6	3	0.135	-
	2	4.6	4	0.328	.445	3.2	1	0.075	-	1.5	3	0.693	-
	3	8.1	4	0.087	.403	0.2	1	0.677	-	8	3	0.047	-
	4	5.2	4	0.266	.403	0.2	1	0.664	-	5	3	0.169	-
	5	3.8	4	0.431	.512	0.6	1	0.425	-	3.2	3	0.365	-
	6	6	4	0.202	.403	0.7	1	0.406	-	5.3	3	0.152	-
	7	2.1	4	0.718	.758	1	1	0.311	-	1.1	3	0.785	-
	8	5.2	4	0.271	.403	2.8	1	0.093	-	2.3	3	0.505	-
	9	0.9	4	0.924	.924	0	1	0.994	-	0.9	3	0.825	-
	10	4	4	0.411	.512	1.6	1	0.212	-	2.4	3	0.493	-
	11	6.7	4	0.151	.403	6	1	0.015	-	0.8	3	0.857	-
	12	6.8	4	0.149	.403	3.9	1	0.049	-	2.9	3	0.413	-
	13	2.6	4	0.625	.699	0.3	1	0.558	-	2.3	3	0.519	-
	14	9.4	4	0.053	.348	3.3	1	0.071	-	6.1	3	0.107	-
	15	12.9	4	0.012	.228	2.2	1	0.141	-	10.7	3	0.013	-
	16	5.1	4	0.276	.403	4.1	1	0.044	-	1.1	3	0.785	-
	17	6.8	4	0.145	.403	2.5	1	0.114	-	4.3	3	0.231	-
	18	9.3	4	0.055	.348	2.9	1	0.090	-	6.4	3	0.095	-
	19	6.2	4	0.185	.403	1.3	1	0.264	-	4.9	3	0.178	-

Note. B. H. p = Benjamini-Hochberg correction value.

±.60 logits across all scales/subscales. Marginal reliability varied slightly over the intervals, but was generally high on all general factor scales (see Table 4.27). These were vast improvements to measurements from the prior round.

Validity assessment results. Since the second round was only large enough for one calibration (i.e., not enough for a development and validation sample), many of validity assessments were not possible to perform Across all of the scales and subscales in round 2, there were only two outlier cases, both of which occurred in the social cognitive (knowing) scale (z = 3.47, 3.61). After removing the two outlier scores, non-factorial design MANOVA analyses were then conducted over race, present religion, gender, classification, and STEM major across all three scales (economic, environmental, and social). However, in round 2, there were few respondents in some of the STEM major categories (ranging from 3 to 8 in the lowest categories). To prevent this from

Table 4.27
Round 2 Standard Error of Estimate (SEE) and Marginal Reliability Estimates for STEMSEI Scales

		SEE	Marginal Reliability and Corresponding Interval		
			[-3, 3]	[-2, 2]	[-1, 1]
Economic Scale	General Factor	.38	.91	.92	.93
	Emotional Subscale	.50	.67	.68	.67
	Cognitive (Knowing) Subscale	.46	.79	.81	.83
	Cognitive (Solving) Subscale	.49	.76	.81	.84
Environmental Scale	General Factor	.55	.89	.91	.91
	Emotional Subscale	.54	.74	.77	.80
	Cognitive (Knowing) Subscale	.60	.76	.78	.80
	Cognitive (Solving) Subscale	.55	.77	.82	.86
Social Scale	General Factor	.49	.88	.89	.90
	Emotional Subscale	.58	.59	.65	.70
	Cognitive (Knowing) Subscale	.58	.71	.73	.75
	Cognitive (Solving) Subscale	.54	.73	.78	.82

skewing the results of the MANOVA, respondents who indicated a STEM major of chemistry, computer science, mathematical sciences, physics and astronomy, psychology, or STEM education were removed. This made the number of cases per cell exceed the number of dependent variables being assessed, as recommended by Meyers et al. (2006, p. 375).

After compensating for type I errors by applying a Benjamini-Hochberg correction, the non-factorial design MANOVA analyses found statistically significant differences along STEMSEI scores from round 1 (see Table 4.28). Statistically significant differences were found between groups in gender and STEM major along the economic scales (see Table 4.28). Statistically significant differences were found across

Table 4.28
Non-factorial Design MANOVA Analyses for STEMSEI Scale Scores

		F	*df*	*p*	*B. H. p**	η^2
Economic Scales	Race	1.775	8, 488	.080	.107	.028
	Gender	12.380	4, 228	**< .001**	**.002**	**.178**
	Present Religion	.878	4, 178	.478	.478	.019
	Classification	.715	4, 231	.582	.582	.012
	STEM major	5.193	40, 818	**< .001**	**.002**	**.203**
Environmental Scales	Race	2.604	8, 488	**.009**	**.011**	**.041**
	Gender	10.771	4, 228	**< .001**	**.002**	**.159**
	Present Religion	6.131	4, 178	**< .001**	**.002**	**.121**
	Classification	2.884	4, 231	**.023**	**.023**	**.048**
	STEM major	5.456	40, 818	**< .001**	**.002**	**.211**
Social Scales**	Race	5.502	4, 228	**< .001**	**.002**	**.088**
	Gender	5.395	4, 228	**< .001**	**.002**	**.086**
	Present Religion	2.948	4, 178	**.022**	**.022**	**.062**
	Classification	3.780	4, 231	**.005**	**.006**	**.061**
	STEM major	2.648	40, 818	**< .001**	**.002**	**.115**

Note. * - Values for the Benjamini-Hochberg correction.
** - Scores from the social scales were not included in these analyses due to non-normality.

all demographic variables on the environmental scores, as expected (see Table 4.28). Statistically significant differences were found across all demographic variables on the social scores as well (see Table 4.28). Post-hoc ANOVA tests for these results (assessed at the α = .01 level) revealed various statistically significant differences along race, gender, present religion, and STEM major across the economic, environmental, and social scales (see Table 4.29). Post-hoc t-tests on the statistically significant results of the ANOVA analyses revealed several statistically significant differences across the various scales for different demographic variables (see Table 4.30).

Table 4.29
Post-hoc ANOVA Analyses for STEMSEI Scale Scores with Benjamini-Hochberg correction

		F	*df*	*p*	*B. H. p**	η^2
Gender	General Economic	24.646	1, 232	**< .001**	**.002**	**.096**
	Emotional Economic	.155	1, 232	.694	.694	.001
	Cognitive (know) Economic	22.696	1, 232	**< .001**	**.002**	**.089**
	Cognitive (solve) Economic	.606	1, 232	.437	.583	.003
Race	General Environmental	.016	1, 232	.901	.901	.000
	Emotional Environmental	5.538	1, 232	.019	.076	.023
	Cognitive (know) Environmental	.743	1, 232	.390	.780	.003
	Cognitive (solve) Environmental	.040	1, 232	.842	.901	.000
Gender	General Environmental	7.198	1, 232	.008	.011	.030
	Emotional Environmental	17.347	1, 232	**< .001**	**.002**	**.070**
	Cognitive (know) Environmental	26.800	1, 232	**< .001**	**.002**	**.104**
	Cognitive (solve) Environmental	6.679	1, 232	**.010**	**.010**	**.028**
Present Religion	General Environmental	11.019	1, 182	**.001**	**.002**	**.057**
	Emotional Environmental	15.540	1, 182	**< .001**	**.002**	**.079**
	Cognitive (know) Environmental	.760	1, 182	.384	.462	.004
	Cognitive (solve) Environmental	.543	1, 182	.462	.462	.003
Classification	General Environmental	1.297	1, 235	.256	.341	.006
	Emotional Environmental	.254	1, 235	.615	.615	.001
	Cognitive (know) Environmental	3.029	1, 235	.083	.166	.013
	Cognitive (solve) Environmental	6.635	1, 235	.011	.044	.028

Table 4.29 (cont.)
Post-hoc ANOVA Analyses for STEMSEI Scale Scores with Benjamini-Hochberg correction

		F	*df*	*p*	*B. H. p**	η^2
Race	General Social	13.405	1, 231	**< .001**	**.004**	**.055**
	Emotional Social	6.367	1, 231	.012	.024	.027
	Cognitive (know) Social	1.693	1, 231	.195	.195	.007
	Cognitive (solve) Social	2.426	1, 231	.121	.161	.010
Gender	General Social	.869	1, 231	.352	.352	.004
	Emotional Social	8.004	1, 231	.005	.014	.033
	Cognitive (know) Social	3.343	1, 231	.069	.092	.014
	Cognitive (solve) Social	7.476	1, 231	.007	.014	.031
Present Religion	General Social	1.878	1, 181	.172	.344	.010
	Emotional Social	9.236	1, 181	.003	.012	.049
	Cognitive (know) Social	1.240	1, 181	.267	.356	.007
	Cognitive (solve) Social	.163	1, 181	.687	.687	.001
Classification	General Social	.026	1, 234	.872	.872	.000
	Emotional Social	8.182	1, 234	.005	.020	.034
	Cognitive (know) Social	5.282	1, 234	.022	.044	.022
	Cognitive (solve) Social	1.047	1, 234	.307	.409	.004

Note. * - Values for the Benjamini-Hochberg correction.

Table 4.30
Post-hoc t-test Analyses for STEMSEI Scale Scores Identifying Sub-group with Higher Average Scores and Cohen's D

	Scale	*t*	*df*	*p*	*Higher Average Scores*	*Means and Stan. Dev.*	*Cohen's d*
Gender	General Economic	-4.964	**231**	**< .001**	**Males**	M = .31, SD = .91 M = -.29, SD = .92	**.656**
	Cognitive (know) Economic	-4.764	**213**	**< .001**	**Males**	M = .26, SD = .91 M = -.29, SD = .85	**.625**
	Emotional Environmental	4.165	**231**	**< .001**	**Females**	M = .27, SD = .76 M = -.17, SD = .86	**.542**
	Cognitive (know) Environmental	-5.177	**231**	**< .001**	**Males**	M = .25, SD = .75 M = -.26, SD = .77	**.671**
	Cognitive (solve) Environmental	-2.584	**231**	**.010**	**Males**	M = .15, SD = .84 M = -.14, SD = .91	**.331**
Race	General Social	3.661	**231**	**< .001**	**Students of Color**	M = .34, SD = .81 M = -.08, SD = .81	**.519**
Present Religion	General Environmental	3.319	**181**	**.001**	**Non-Christians**	M = .24, SD = .82 M = -.16, SD = .77	**.503**
	Emotional Environmental	3.942	**181**	**< .001**	**Non-Christian**	M = .31, SD = .85 M = -.18, SD = .78	**.601**

Correlations among scales and subscales. When considering the correlations between the various scales and subscales of the STEMSEI, there were expected and unexpected results. First, all corresponding subscales (emotional, cognitive (knowing), and cognitive (solving)) correlated positively and statistically significantly (see Table 4.31), with Pearson's *r* ranging from .182 to .462. This meant that as emotional sustainability engagement increased in one sustainability domain, it increased, in general, on the other two. On the other hand, only one pair of the general scales did not correlate positively and statistically significantly: the economic and social general scales. This indicates one level of internal consistency within the scores produced by the STEMSEI.

Finally, several correlations of STEMSEI scores with age were statistically significant, including; the general environmental sustainability engagement score, $r(223) = .172$, $p = .010$; the emotional social sustainability engagement score, $r(223) = .131$, $p = .049$; the cognitive (knowing) score across the economic, $r(223) = .213$, $p = .001$, environmental, $r(223) = .186$, $p = .005$, and social, $r(223) = .172$, $p = .010$ scores; the

cognitive (solving) environmental score, $r(223) = -.173$, $p = .009$. Except for the cognitive (solving) environmental scale scores, these correlations indicate that as age increases, scores on these scales, in general, increase as well.

Table 4.31
Pearson's Correlations Among STEMSEI Scores

		General			Emotional			Cognitive (know)			Cognitive (solve)		
		Econ	Env	Soc	Econ	Env	Soc	Econ	Env	Soc	Econ	Env	Soc
General	Econ	1											
	Env	.15*	1										
	Soc	0.05	.35**	1									
Emotional	Econ	-0.05	.13*	.14*	1								
	Env	-.27**	0.12	.39**	.18**	1							
	Soc	-0.05	0.05	0.11	.22**	.46**	1						
Cognitive (know)	Econ	0.04	.35**	.28**	0.09	-.16*	-.19**	1					
	Env	.18**	.34**	-0.06	-0.04	-.35**	-0.12	.38**	1				
	Soc	0.05	.41**	.16**	0.05	-0.06	-0.02	.44**	.36**	1			
Cognitive (solve)	Econ	0.05	0.06	.34**	0.03	0.01	-0.06	.17**	-0.08	-0.01	1		
	Env	.46**	.23**	-0.05	0.03	-.23**	-0.06	-0.04	0.02	-0.09	.32**	1	
	Soc	.32**	-.13*	.17**	-0.11	-0.08	-0.09	-0.03	-0.03	-0.06	.24**	.31**	1

Note. Econ = Economic Scale; Env = Environmental Scale; Soc = Social Scale; N = 253 for all correlations.
* $p < .05$
** $p < .01$

Copyright © David L. Little II 2014

Chapter V

Discussion, Implications, and Conclusions

This chapter first discusses the results of this study in context of sustainability engagement in post-secondary STEM students. First, each of the hypotheses for the STEMSEI are discussed with respect to the results. Second, issues of concern in the development, evaluation, and validation of the STEMSEI is discussed. Third, areas where the STEMSEI might be refined to improve instrument functioning are considered. With those issues of concern and future improvements in mind, data trends that inform current sustainability engagement in post-secondary STEM students are then discussed. Results are then synthesized to form final conclusions. Lastly, looking to future research, implications for measuring sustainability engagement and other sustainability education outcomes in post-secondary STEM students are discussed.

Hypotheses Testing

This section details the results of the hypotheses testing based on the results presented in Chapter IV. Each research question is discussed separately and the associated hypotheses are individually addressed.

Research question 1. The first research question ("To what extent can a convergent theoretical framework for sustainability engagement in post-secondary STEM students be achieved between sustainability experts across the STEM disciplines?") was connected to phase 1 of the study and sought to establish the theoretical framework of the STEMSEI. Despite claims in the literature that generalizable sustainability instruments are not possible (Shriberg, 2002), the hypothesis for this research question was that a convergent theoretical framework was possible. Results from phase 1 indicate this was true since the content experts were able to understand the *sustainability engagement framework* and the *quantifiable definition of sustainability*. Results from phase 2 also support this conclusion since the alignment study with the items of the STEMSEI were able to be appropriately aligned independently by content experts to the appropriate aspects of the *sustainability engagement framework* and the *quantifiable definition of sustainability*. Phase 3 results also support this since the factor analysis results indicated an interpretable and predicted factor structure of the instrument.

Research question 2. Phase 2 focused on a different research question ("To what extent can items that measure unique types of sustainability engagement and can be interpreted by post-secondary sustainability educators and post-secondary STEM students be developed across the STEM disciplines?"). The hypothesis to this research question was that the factor structure of the responses to the STEMSEI could distinguish along either sustainability domain or engagement type. Similar to the results concerning the previous research question, the alignment study, cognitive interviews, and factor analysis techniques all point to the same conclusion: that items that align to unique types of engagement can be interpreted by both post-secondary sustainability educators and STEM students. Else, the items of the STEMSEI would not have been aligned appropriately by a majority of the content experts, nor would the items have been understood during the student cognitive interviews. Finally, the factor pattern matrices (see Table 4.5 and 4.18) would not have aligned to the structure of the *sustainability engagement framework* (i.e., along emotional, cognitive (know), and cognitive (solve) engagement).

Research question 3. The final research question ("To what extent does the STEMSEI produce interpretable and useable/meaningful results with respect to sustainability education at the post-secondary level?") had three separate hypotheses. They are discussed individually below.

Hypothesis 1. The statistically significant differences in the environmental engagement measures were not present with respect to race, counter to expectations (see Table 4.29). On the other hand, statistically significant differences in environmental engagement were found with respect to gender and religious beliefs (see Table 4.30). Moreover, the direction of these significant differences were as predicted by gender (see Johnson, Bowker, and Cordell, 2004; Olli, Grendstad, & Wollebaek, 2001) and religious beliefs (Lalonde & Jackson, 2002; Thapa, 1999). In addition, females were more engaged in emotional environmental sustainability engagement while males were more engaged in both forms of cognitive environmental sustainability engagement. This reflects the findings from Johnson et al. (2004) and Olli et al. (2001) that females are more engaged with environmental issues, which, with respect to the STEMSEI, could be interpreted as emotional engagement. On the other hand, Johnson et al. (2004) and Olli

et al. (2001) found males more engaged in environmental activism, which could be interpreted as either cognitive or behavioral engagement. Finally, environmental engagement did also vary statistically significantly based on religious beliefs (see Table 4.30). Moreover, these differences were as predicted by Lalonde and Jackson (2002) and Thapa (1999), with Christians scoring statistically significantly lower than non-Christians.

Hypothesis 2. Statistically significant differences were found across all STEM majors in all STEMSEI scores (see Table 4.28). Again, these differences were expected due to the content knowledge difference between different STEM majors (i.e., cognitive engagement) and also the proclivity natural infusion of certain opinions/preferences that may accompany some STEM majors (i.e., emotional engagement). The STEMSEI delineates sustainability along content (i.e., economic, environmental, and social domains), and the engagement various STEM majors have with sustainability varies statistically significantly based on major.

Hypothesis 3. Also as predicted, there were no statistically significant differences in sustainability engagement based on classification (see Table 4.29). No significant differences were expected here since graduate coursework would not theoretically make post-secondary STEM students any more engaged in sustainability than undergraduate coursework.

Unexpected result. It was not expected that students of color would be statistically significantly more engaged in general social sustainability than whites (see Table 4.30). However, this result may indicate the influence of minority status on the perceptions of social sustainability issues. This should be explored further in the future.

Hypothesis testing conclusions with respect to validity. With the exception of significant differences based on race in environmental sustainability engagement measures, all hypotheses were confirmed in the results. This lends validity evidence to the scores produced by the STEMSEI. Follow-up research should test the validity of this finding: does environmental sustainability engagement statistically significantly differ based on race? One explanation for this result being counter to those predicted by Johnson et al. (2004) and Olli et al. (2001) could be the age group difference surveyed in

this study and those studies (i.e., post-secondary students as compared to high school students).

Moreover, the unexpected result of significant differences in social sustainability engagement scores due to race could be indicative of a lack of validity in the STEMSEI. However, for this to be confirmed, further validity testing is required in the population.

Overall, the validity evidence for the STEMSEI seems strong, and results from the factor analyses and the IRT analyses both support the validity of the STEMSEI as well (i.e., the confirmation of the hypotheses from research questions 1 and 2). While a complete validation phase could not be completed as recommended by Benson and Clark (1982), these preliminary validity tests do help establish some benchmarks by which to evaluate the validity of the STEMSEI.

Considerations in STEMSEI Development, Evaluation, and Validation

While this study has shed light on how sustainability engagement exists among post-secondary STEM majors, the size and scope of this study left areas for improvement. Given that sustainability involves problems that are large, complex, and require perspectives from multiple disciplines to solve (Meadows, 2008) and engineers have found sustainable measures and practices too difficult to incorporate into their work (Brown, 2011), it was unsure if post-secondary STEM students would feel inclined to participate in a study focusing on sustainability. While low response rates were seen as a possibility and a possible threat to validity, the data in both rounds provided a spread across all response options for all items. This has been shown to be one of the most important factors when estimating item parameters for multidimensional graded response models (de Ayala, 1994). Hence, the low response rate was in part compensated for by the spread of responses across all options. However, this could also mean that future calibrations of the STEMSEI estimate different item parameters than those presented in Tables 4.9, 4.10, and 4.21 to 4.23. Invariance of item parameters should be evaluated to determine the validity of the STEMSEI across the population.

The content experts had feedback concerning the blind review process implemented in round 1 of step 2b of phase 1 (see Figure 3.1) versus the non-blind review implemented in round 2 of step 2b of phase 1. Consistently, the content experts liked the second system better, as they felt it was easier to respond to and more efficient.

In the case of items that were multidimensional with respect to sustainability domains, the methodology adapted from Norcini et al. (1993) used in the second round also allowed an easier way to indicate proportionality of importance to the various sustainability domains. While the first round was blind to remove any bias from the researcher, this may be a factor to consider in similar studies, as it may be an unneeded precaution as the STEMSEI is further developed. Moreover, it is the researcher's expectation that the methodology used in phase 1 could easily be replicated to produce other similar constructs of interest in sustainability education.

The factor analysis results seem to indicate that the sustainability engagement framework is appropriate for conceptualizing sustainability engagement in post-secondary STEM students, at least when considering emotional and cognitive engagement. However, the invariance of the factor structure of the STEMSEI, or lack there-of, should be explored further with other samples. Specifically, the factor structure of sustainability engagement across STEM majors should be compared for possible differences. Given that sustainability is a topic gaining traction even in general engineering courses, there could be differences of the factor structure of the STEMSEI even within this population.

Moreover, the responses to the interviews conducted in phase 3 seem to indicate that post-secondary STEM students see a connection between the economic, environmental, and social domains of sustainability, but many are unable to articulate what that connection is or describe it in detail. This could explain why the EFA's along engagement type did not perform as well as extractions along the domains of sustainability. This is corroborated in the qualitative data since STEM students were, in general, able to describe how each domain separately connected to sustainability. It may be that there is too much noise in the data when considering items across the three domains of sustainability in one scale (e.g., along engagement type).

While the unidimensional differential item functioning (DIF) tests from round 2 did not reveal any items functioning differently over groups, multidimensional DIF analyses should be considered since they are more appropriate given the multidimensionality of the bifactor IRT model that was used. Future research should identify if bias is present in the latter version of the STEMSEI (see Appendix I). If bias

in an item is identified, it should be contextualized within the population to understand the cause of that bias.

Also, it is of particular interest that in both factor analyses and IRT analyses, the non-orthogonal bifactor model consistently fit the STEMSEI scales better than any other factor analysis model and IRT model. It is likely that the simultaneous unidimensionality and multidimensionality of sustainability engagement across economic, environmental, and social contexts makes bifactor models (both factor analysis and IRT versions) ideal for estimating latent trait scores for these constructs. This is especially convenient for post-secondary sustainability educators who use the instrument since use of the general scale scores may be preferable over subscale scores. The superior fit of bifactor models should be tested in future samples to determine the generalizability of this result.

Moreover, the ability of the STEMSEI to predict statistically significant differences across STEM majors (see Table 4.28) offers domain evidence for the instrument. The large effect sizes indicate there are some STEM majors that have a large and distinct advantage over other STEM majors (see Table 4.28). These differences should be more clearly understood so that we can understand how we might tailor sustainability education efforts to suit the needs of the different STEM majors.

Finally, while several validity checks did seem to indicate quantitative integrity in the measures produced by the STEMSEI, this is a new instrument that needs to be further refined and edited. When considering that sustainability is still an evolving topic on educational and administrative levels at post-secondary institutions (Christensen, Thrane, Jorgensen, & Lehmann, 2009; Wright, Ironside, & Swynn-Jones, 2009), this means that sustainability education will necessarily change and so must the STEMSEI if it is to continue to measure a changing construct.

Refining the STEMSEI

Though the results of the functioning of the STEMSEI were strong on both the item and instrument level, the STEMSEI can be further refined to increase measurement accuracy. More items that align in both engagement type and sustainability domain should be written. First, since the item stems currently employed across the STEMSEI exhibit homogeneity, there could be inflated correlations between items, factors, scores, and/or subscores. Since the primary focus in each item was item content (i.e., not the

item stem), this should be minimized within the STEMSEI. Moreover, if item homogeneity was strong within the STEMSEI, the EFA's along engagement type would have been stronger than those across sustainability domain. Second, more effective and efficient items could be produced that would make person location estimates more accurate or could lead to a short-form of the STEMSEI to be developed.

Due to the evolution of sustainability and the focus it is gaining in post-secondary institutions, the factor structure of the STEMSEI will likely be influenced by these factors as students become more involved in sustainability efforts and programs. Thus, the factor structure of the STEMSEI will not likely be static and should be reevaluated and refined if necessary. Currently, both qualitative and quantitative data supports the idea that post-secondary STEM students do not see the economic, environmental, or social domains of sustainability as a cohesive concept (i.e., that sustainability is unidimensional). On the contrary, they see them as more distinct, with overlap occurring only when societal structures necessitate it (i.e., the environment being our only source of resources for product manufacturing, stressing the importance of economics over environmental concerns, etc.). This presents methodological concerns in future analyses with the STEMSEI. If implemented with a sample of post-secondary STEM students that are very engaged in sustainability (both emotionally and cognitively), then the factor structure of the responses to the STEMSEI may indeed be unidimensional, a possibility suggested by the engineering content experts in phase 1. Researchers need to bear in mind that certain measurement properties are constituted by the target sample, not the instrument used, and that these properties need to be reassessed in some samples.

The STEMSEI should be able to discriminate across all STEM majors, which was one of the instrument’s objectives. This discrimination power was evidenced in the instrument in both rounds. However, this should be tested with a larger sample across all STEM majors. Moreover, the effect size should theoretically be strong across all of the different scores generated by the STEMSEI. This is because disciplinary differences will theoretically promote high engagement when disciplinary content overlaps with sustainability topics. However, since the senior engineering student from phase 2 was able to identify sustainability as a cohesive unit of economic, environmental, and social issues, it may be that some majors are being instructed in ways that minimize differences

across STEMSEI scores. Ideally, this is the type of instruction that sustainability related courses should aim for, as post-secondary students need this type of instruction to learn the skills and perspectives they need to face the complex problems of sustainability (Bencze, Sperling, & Carter, 2011; Erdogan, 2010).

The largest refinement of the STEMSEI is the need for a portion of the instrument that measures behavioral sustainability engagement. This is crucial since the connection between knowledge for and action for sustainability have been absent across the literature (see Christensen, Thrane, Jorgensen, & Lehmann, 2009; Davis, Edmister, Sullivan, & West, 2003; Hodson, 2003; O'Connell, Potter, Curthoys, Dyment, & Cuthbertson, 2005; Thapa, 1999; Van Kerkhoff & Lebel, 2006; Wright, Ironside, & Gwynn-Jones, 2009). The development, evaluation, and validation of this new portion of the instrument may or may not be similar to methodologies employed here for emotional and cognitive sustainability engagement. Regardless, this is an important and major undertaking for the next steps of the STEMSEI.

Implications for Sustainability Engagement in Post-secondary STEM Students

The STEMSEI proved to be a more complex instrument than originally expected. However, from this complexity came information on the sustainability engagement of post-secondary STEM students from multiple new perspectives. The STEMSEI is able to measure sustainability engagement in the population along the delineation of economic, environmental, and social sustainability, which is important for sustainability education efforts. However, the STEMSEI can also contextualize those measures in ways some quantitative instruments cannot. The subscale scores (emotional, cognitive (knowing), and cognitive (solving)) add a new dimension to the meaningfulness of these scores. It will allow researchers to contextualize results in many different ways that could further improve the impact of scores generated by the STEMSEI. Some of these contextualizations are discussed further next.

Differences in post-secondary STEM student sustainability engagement. The differences across various subsamples of post-secondary STEM students highlight some of the challenges to sustainability education. This study focused on differences related to race, gender, current religious practices, and classification. Additional differences also analyzed differences along STEM majors. While the results of the analyses of these

differences provide powerful insight, recall that a limited sample was used to generate these findings, thus limiting generalizability. None-the-less, the findings of this study break new ground on possible differences in sustainability engagement that need to be researched further. These differences are discussed further here.

Recall that in round 1 DIF among environmental items with respect to race was detected. The fact that DIF among these items was not detected in round 2 may indicate a false positive for the first round. However, consider the DIF results from round 1. Item locations for white students were substantially less than item locations for students of color. This is similar to differences Johnson, Bowker, and Cordell (2004) found in environmental belief and behavior based on ethnic variation. Whites were found to score higher on the New Environmental Paradigm (see Dunlap, Van Liere, Mertig, & Jones, 2000) when compared to blacks and foreign-born Latinos. It could be that DIF on environmental items along race means that white students respond favorably to these items because of self-perceived social expectations. On the other hand, students of color may not experience social expectations to support certain environmental topics/issues. This phenomenon requires further research to understand cultural impacts on sustainability engagement.

Concerning the gender DIF identified in environmental items in round 1 (but not round 2), there are two ways of looking at this. First, recall that estimated item locations for females were less than those estimated for males on many of the environmental items. Moreover, statistically significant differences between emotional and cognitive (knowing) environmental engagement were found between females and males. These results were expected, and are similar to those found by Johnson, Bowker, and Cordell (2004) and Olli, Grendstad, and Wollebaek (2001). This means that females and males responding equivalently do not indicate the same level of environmental sustainability engagement. This could mean that females are more apt to indicate a favorable response to these to environmental topics/issues because of social norms (e.g., higher scores on emotional environmental sustainability engagement). If this is the case, it is possible that female emotional environmental sustainability engagement could be leveraged to encourage females to enter the STEM pipeline, which is a pressing issue (Hanson, 2004; National Science Board, 2010; Smeding, 2012). On the other hand, the fact that males

scored higher on cognitive (knowing) environmental sustainability engagement might reflect the fact that disciplines that study these issues are male-dominated. This research is inadequate to fully address these hypotheses, however, and should be considered for further study. It may be that these results are not replicated in future applications of the STEMSEI because the results from round 1 may have been false positives.

The differences in Christians and non-Christians with respect to their environmental sustainability engagement is of particular interest for future research. The fact that one of the STEM education students, who is also a Christian, highlighted this very distinction in the phase 3 interviews (see Appendix F) triangulates (Creswell, Plano Clark, Gutmann, & Hanson, 2003; Erzberger & Kelle, 2003) this finding between the qualitative and quantitative data. The context of this distinction has been attributed to certain faith-based perspectives (e.g., the Earth has "infinite resources" and "infinite space" to accommodate humans; see Lalonde & Jackson, 2002; Thapa, 1999) that seem to diminish environmental concern or engagement. The item content of the New Environmental Paradigm instrument (see Dunlap, Van Liere, Mertig, & Jones, 2000) were written, in part, to reflect environmental limits to growth and anti-anthropocentrism (from environmental/ecological contexts; Dunlap, Van Liere, Mertig, & Jones, 2000). These areas draw upon faith-based perspectives that are counter to environmental sustainability by their very nature (e.g., infinite resources and space). While mixing science and faith can be a volatile social issue, discourse on these issues must take place in post-secondary environments. A mutually compatible resolution to these seemingly competing perspectives is possible, as evidenced by the Christian STEM education student who has bridged her faith-based perspectives with her scientific perspectives. Future research should identify paths in which individuals reach such mutually compatible resolutions.

The potential implications for differences in sustainability engagement along STEM majors are possibly the most interesting and important with respect to sustainability itself. As post-secondary programs prepare STEM students to face the sustainability problems of the future it almost necessitates that differences diminish amongst STEM majors prepared to face sustainability issues. This could be one evaluative test for sustainability education programs to ensure graduates do grow in

economic, environmental, and social sustainability outcomes (e.g., sustainability engagement). Even if significant differences persist, the effect size of such differences should be negligible.

Sustainability engagement invariance across post-secondary STEM students. The tenability of IRT model invariance across samples (i.e., the invariance of item parameters estimated for the STEMSEI across different samples) also has implications for post-secondary sustainability education efforts. For example, if model invariance were found, this would prove that generalizable models for sustainability are possible, counter to Shriberg's (2002) claim. However, it may simply be that generalizable models for sustainability education were outside the perspective of what Shriberg (2002) was considering. Regardless, if a generalizable model for sustainability engagement is possible and average measures of sustainability engagement can be established across multiple levels (i.e., by state, region of country, national level, etc.), this would give sustainability educators a powerful platform to begin evaluating and discussing sustainability pedagogy and curricular options with respect to sustainability engagement. However, if a generalizable model that is accessible to faculty members across STEM is possible for sustainability engagement, then generalizable models for other important sustainability education issues also exist (i.e., transdisciplinary effects on sustainability education outcomes).

With respect to the validity of the STEMSEI, there should be a large amount of invariance across the instrument. These areas include (1) item parameters estimated across different samples, (2) person location estimates obtained from item parameters estimated from different samples, (3) a small root mean square difference between corresponding option response functions of two models, and (4) empirical model-fit when scoring a sample using item parameters from a different sample. These validity checks help ensure that the STEMSEI is functioning equivalently across the population. However, if one of these validity checks fails, it does not necessarily mean that the STEMSEI is not functioning as intended. It could be that there are particular properties of the sample being assessed that change how the STEMSEI functions within that sample (i.e., number of sustainability-related courses). Researchers must be ever vigilant to validate the instruments used within a particular sample and cannot assume invariance of

the STEMSEI across post-secondary STEM students; invariance needs to be tested and affirmed for each sample.

Levels of sustainability engagement across STEM majors. The ability of the STEMSEI to measure sustainability engagement across multiple contexts (sustainability domain and engagement type) is powerful to begin understanding how these different types of engagement interact with one another. The correlations in Table 4.31 begin to provide insight into these interactions. Positive correlations along each engagement type (i.e., emotional environmental engagement and emotional economic engagement) indicate that as an engagement type in one domain of sustainability increases, engagement of the same type in other domains of sustainability increase. From the perspective of sustainability curricula, this should be expected if students are to view the domains of sustainability as unified. However, what is unexpected is how different engagement types interact.

For example, as general economic engagement increases, emotional environmental engagement decreases. This result is to be expected since economic concerns are often seen as more important than environmental issues (Ameer & Othman, 2012; Anderson, 2009; Dyllick & Hockerts, 2002; Fairbrass & Zueva-Owens, 2011; Johnston, Everard, Santillo, & Robèrt, 2007; Norman & MacDonald, 2004). This is even further contextualized by in cognitive (knowing) economic engagement; as students increase their knowledge of how to engage in economic sustainability, they, on average, decrease in emotional engagement with both environmental and social sustainability. These correlations all suggest a similar message. However, as general economic engagement increased, both types of cognitive environmental engagement (knowing and solving) also increased on average, leading to a somewhat counter result. It may be that as economic engagement reaches a certain level, the importance of the other domains of sustainability become more apparent. It could also mean that knowledge for solving environmental sustainability problems could increase while emotional environmental sustainability engagement remains unaffected. Results as this demonstrate the importance and effectiveness of sustainability education efforts at the post-secondary level.

Looking to the future when behavioral engagement may be added to the STEMSEI, it is important to note that many of the STEM students interviewed in phase 3 indicated that they would not want to work in fields related to sustainability. However, each student acknowledged the importance of work in such fields. While this statement is indicative of behavioral engagement, there are also implications about the emotional engagement of participants in this notion. Consider the many negative correlations between the emotional scales and the two cognitive scales (see Table 4.31). Of these, 13 out of 18 of these correlations were negative, though only 4 of these correlations were statistically significant (see Table 4.31). This echoes the results Brown (2011) found engineers found it hard to incorporate sustainability in their practices even though they saw the importance of it.

Finally, the sustainability engagement framework theorizes that behavioral engagement cannot precede cognitive engagement in sustainability. That is, you must know how engage (knowing) and how to solve problems (solving) in sustainability before you can actually do so. However, even before this, one must understand how concepts from their major can be applied to sustainability problems (knowing) before they can attempt to solve those problems cognitively with disciplinary expertise (solving). This means that as cognitive (knowing) engagement increases, so should cognitive (solving) engagement. In general, this is unfortunately not the case. Correlations between these types of engagement were close to 0 except for the cognitive economic engagement types (see Table 4.31). Studies have revealed that the majority of post-secondary students think environmental problems "should be addressed by Environmental Scientists and that an environmental literacy requirement will constitute an additional academic burden on them" (Aighewi & Osaigbovo, 2009, p. 632). Hence, it is no surprise that STEM majors, in general, do not realize the capacity of their major to know and solve environmental sustainability problems. Moreover, the fact that only economic cognitive engagement types correlate reflects the fears of Bencze and Carter (2011) and Hodson (2003). Namely, post-secondary institutions may be reluctant to fully commit to sustainability education since that would require revising cultural perceptions the utility of science (i.e., to be profitable; Bencze & Carter, 2011; Hodson, 2003).

Future Uses of the STEMSEI

The STEMSEI itself was designed for two main purposes. First, it was designed to give sustainability educators an appropriate tool to evaluate sustainability engagement in post-secondary environments. Instrumentation employed for related studies has been outdated and does not accurately measure the constructs the researcher purported to be researching (see Mann, Harraway, Broughton-Ansin, Deaker, & Shephard, 2013; Schneiderman & Freihoefer, 2012). The STEMSEI fills this research gap by providing researchers with an instrument to measure sustainability engagement, which should arguably be an educational outcome for all sustainability-related courses. It is intended for use to measure sustainability engagement outcomes associated with any treatment that should increase sustainability engagement (i.e., a course in sustainability). The STEMSEI is not intended to measure students' knowledge of sustainability or to evaluate teaching in sustainability-related courses. It should only be used to evaluate sustainability curricula if sustainability engagement is an educational outcome of the curricula; no extensions beyond the curricula's impact on sustainability engagement should be extrapolated from the STEMSEI or the scores it produces.

The STEMSEI is an extremely adaptable instrument with respect to this first use. The quantifiable definition of sustainability and the Sustainability Engagement Instrument are foundational frameworks for the STEMSEI that allow the instrument to adapt to the incremental changes that will inevitably occur in sustainability education. Moreover, the STEMSEI offers variety of ways in which a researcher might measure sustainability engagement (i.e., measuring only emotional sustainability engagement, measuring only social sustainability engagement, measuring only general sustainability engagement along each sustainability domain, etc.). Hence, the timeliness of this instrument cannot be understated, nor can its versatility in measuring sustainability engagement in the population.

The second purpose of the STEMSEI was to give sustainability educators across the United States (and possibly the globe) a tool to compare and contrast sustainability engagement in various samples/subpopulations. This is the first step in identifying effective and efficient pedagogical and curricular elements of sustainability education programs. Though the need for improvements to sustainability education efforts has been

established (United Nations Educational, Scientific and Cultural Organization, 2005), no instrumentation has surfaced to contribute to a collective pedagogical and curricular view of sustainability education. In fact, many have called for abandoning generalized approaches to sustainability (Kiewiet & Vos, 2007; Shriberg, 2002), which may, in turn, be creating the thought that a generalized approach to sustainability education is also unneeded or problematic. The success of the STEMSEI shows that generalized models for sustainability education outcomes are indeed possible.

Final Implications

While the STEMSEI was designed for two purposes, the success of this study leads to a final purpose of this study. Thorough test development is a demanding process that involves many steps and safeguards to insure proper item and instrument functioning (Downing, 2006). Sustainability engagement is only one sustainability education outcome that should be measured. To that end, the development of future instruments to measure other sustainability education outcomes will be demanding as well. This study can serve as a guide to a methodology to produce such instruments to further advance sustainability education.

Recommendations

A longitudinal study of the functioning of the STEMSEI in post-secondary STEM majors is needed to assess the instruments' invariance across multiple samples. This will also determine the extent to which generalizability of findings can be expected from the STEMSEI.

The need for an instrument that measures behavioral sustainability engagement is crucial to the field. The development of this should commence immediately. Moreover, if possible, the instrument should be designed to be theoretically and methodologically compatible with the STEMSEI so that a complete, unified assessment of sustainability engagement in post-secondary STEM majors can be established.

Finally, trends in sustainability education with respect to sustainability engagement should be researched to identify successful curricular and pedagogical elements. Differences in sustainability engagement rooted in race, gender, current religious practices, and other variables should be identified. Further, the context in which these differences surface in sustainability engagement should be identified and remedied.

Copyright © David L. Little II 2014

APPENDIX A
National Science Foundation STEM Majors

Developed from Gonzalez and Kuenzi (2012) and the National Science Foundation (2006)

CHEMISTRY
Chemical Catalysis
Chemical Measurement and Imaging
Chemical Structure, Dynamics, and Mechanism
Chemical Synthesis
Chemical Theory, Models and Computational Methods
Chemistry of Life Processes
Environmental Chemical Systems
Macromolecular, Supramolecular, and Nanochemistry
Sustainable Chemistry
Chemistry, other (specify)

COMPUTER AND INFORMATION SCIENCE AND ENGINEERING (CISE)
Algorithms and Theoretical Foundations
Communication and Information Theory
Computational Science and Engineering
Computer and Information Security
Computer Architecture
Computer Systems, Networking, and Embedded Systems
Databases
Data Mining and Information Retrieval
Graphics and Visualization
Human Computer Interaction
Informatics
Machine Learning
Natural Language Processing
Robotics and Computer Vision
Software Systems and Software Engineering
CISE, other (specify)

ENGINEERING
Aeronautical and Aerospace Bioengineering
Biomedical
Chemical Engineering
Civil Engineering
Computer Engineering
Electrical and Electronic
Energy
Environmental
Industrial Engineering & Operations Research

Materials
Mechanical
Nuclear
Ocean
Optical Engineering
Polymer
Systems Engineering
Engineering, other (specify)

GEOSCIENCES
Atmospheric Chemistry
Aeronomy
Biogeochemistry
Biological Oceanography
Chemical Oceanography
Climate and Large-Scale Atmospheric Dynamics
Geobiology
Geochemistry
Geodynamics
Geophysics
Glaciology
Hydrology
Magnetospheric Physics
Marine Biology
Marine Geology and Geophysics
Paleoclimate
Paleontology and Paleobiology
Petrology
Physical and Dynamic Meteorology
Physical Oceanography
Sedimentary Geology
Solar Physics
Tectonics
Geosciences, other (specify)

LIFE SCIENCES
Biochemistry
Biophysics
Cell Biology
Developmental Biology
Ecology
Environmental Science
Evolutionary Biology
Genetics
Genomics
Microbiology

Molecular Biology
Neurosciences
Organismal Biology
Physiology
Proteomics
Structural Biology
Systematic Biology
Life Sciences, other (specify)

MATERIALS RESEARCH
Biomaterials
Ceramics
Chemistry of materials
Electronic materials
Materials theory
Metallic materials
Photonic materials
Physics of materials
Polymers
Materials Research, other (specify)

MATHEMATICAL SCIENCES
Algebra, Number Theory, and Combinatorics
Analysis
Applied Mathematics
Biostatistics
Computational and Data-enabled Science
Computational Mathematics
Computational Statistics
Geometric Analysis
Logic or Foundations of Mathematics
Mathematical Biology
Probability
Statistics
Topology
Mathematics, other (specify)

PHYSICS AND ASTRONOMY
Astronomy and Astrophysics
Atomic, Molecular and Optical Physics
Condensed Matter Physics
Nuclear
Particle Physics
Physics of Living Systems
Plasma
Solid State

Theoretical Physics
Physics, other (specify)

PSYCHOLOGY

Cognitive
Cognitive Neuroscience
Computational Psychology
Developmental
Experimental or Comparative
Industrial/Organizational
Neuropsychology
Perception and Psychophysics
Personality and Individual Differences
Physiological
Psycholinguistics
Quantitative
Social
Psychology, other (specify)

SOCIAL SCIENCES

Archaeology
Biological Anthropology
Cultural Anthropology
Anthropology, other
Communications
Decision Making and Risk analysis
Economics (except Business Administration)
Geography
History and Philosophy of Science
International Relations
Law and Social Science
Linguistics
Linguistic Anthropology
Medical Anthropology
Political Science
Public Policy
Science Policy
Sociology (except Social Work)
Urban and Regional Planning
Social Sciences, other (specify)

STEM EDUCATION AND LEARNING RESEARCH

Engineering Education
Mathematics Education
Science Education
Technology Education
STEM Education and Learning Research, other (specify)

APPENDIX B
Phase 1: Content Specialist Semi-Structured Interview Protocol

Instrument Design Review

1. Theoretical Framework Questions
    a. Is it important to measure sustainability engagement in undergraduate and graduate students in STEM?
    b. Will all aspects of sustainability that can be understood by undergraduates and graduates in STEM be covered under the economic, environmental, and social domains? If not, what other domains should be considered?
    c. As a guiding definition for sustainability, do you think the quantifiable definition of sustainability is sufficient for this research? What additions or suggestions would you have for it?
    d. Will all aspects of sustainability engagement be covered under cognitive and emotional engagement? (Acknowledge the obvious limitation of omitting behavioral engagement.)
    e. Is the definition for the cognitive engagement construct sufficient to cover all aspects of engagement with sustainability that relies on factual information pertaining to sustainability or other cognitively driven ideas/interactions? If not, what additions would you recommend?
    f. Is the definition for the emotional engagement construct sufficient to cover all aspects of engagement with sustainability that relies on opinion-based ideas, cultural norms, or other emotionally driven ideas/reactions? If not, what additions would you recommend?
    g. Do you anticipate that students in the sample will be cognitively engaged with sustainability at different levels? Emotionally engaged with sustainability at different levels?
2. Methodology Questions
    a. Will producing a quantitative measure of sustainability in post-secondary students be meaningful? Would such data be useful for you or other content practicioners?
    b. Should economic, environmental, and social items be equally represented? If not, provide a ratio of importance for each.
    c. Should the cognitive engagement subdomains be equally represented? If not, provide a ratio of importance for each.
    d. Should the emotional engagement subdomains be equally represented? If not, provide a ratio of importance for each.

APPENDIX C
Phase 2: Content Specialist Semi-Structured Interview Protocol

Directions - Interview each content expert individually. For each individual item, read questions 1a. through 1c. to the content expert. At the conclusion of item review, ask the content expert question 2.

1. Item Questions – Alignment study
    a. Is the readability of this item suitable for undergraduate and graduate college students? How would you improve readability?
    b. Which type of sustainability engagement (emotional, cognitive, and/or behavioral) does the item align to, if any?
    c. Which domain of sustainability (economic, environmental, and/or social) does the item align to, if any?
2. Other
    a. Are there any other possible sources of bias or improvements you would suggest for this study?
    b. For issues of time, would you recommend omitting any items?
    c. Rank the items in terms of their quality in measuring sustainability engagement.

APPENDIX D
Phase 2: Student Cognitive Interview Protocol

Directions - Read the warm-up exercise to the participant. Help acquaint the participant with the think-aloud process. Then, give the participant the sample question. Finally, use the interview protocol at the bottom to have the participant respond to each item of the instrument.

Warm-up Exercise
First, I am going to do a warm-up with you to acquaint you with what we will be doing for the interview. "Try to visualize the place where you live, and think about how many windows there are in that place. As you count up the windows, tell me what you are seeing and thinking about" (Willis, 1999).

Sample Question
Now I'm going to show you a sample question. <Display the item" it is important to make product manufacturing more efficient for businesses.">

1. As you read this question, tell me what you are thinking about.
2. Can you repeat the question in your own words?
    a. {If misalignment appears to exist, ask this follow-up question.} What was it about the question that made you think of it that way?
    b. {Optional} Is there anything unclear about the question? If so, what is unclear and why?
3. Do you think any of your college peers would interpret the question different from you?
    a. {If they answer yes, ask these follow-up questions.} How do you think they would interpret it different?
    b. Why do you think they would interpret it that way?

Interview Protocol
Now I am going to show you a list of questions pertaining to sustainability. We will go through the same process with all of these items. Your feedback will help me identify items that may be problematic, so please feel comfortable to provide critical feedback. Do you have any questions for me before we start?

<Begin sample items>

1. As you read this question, tell me what you are thinking about.
2. Can you repeat the question in your own words?
    a. { If misalignment appears to exist, ask this follow-up question.} What was it about the question that made you think of it that way?
    b. {Optional} Is there anything unclear about the question? If so, what is unclear and why?
3. Do you think any of your college peers would interpret the question different from you?
    a. {If they answer yes, ask these follow-up questions.} How do you think they would interpret it different?
    b. Why do you think they would interpret it that way?

APPENDIX E
Faculty Liaison Email

Dear <name>:

I am writing you to inquire about including <institution> as part of the sample for my dissertation study. I am a PhD candidate in STEM Education at the University of Kentucky. I am piloting an instrument to measure sustainability engagement in post-secondary STEM (science, technology, engineering, and mathematics) students that will be used to assess curricular outcomes of sustainability-related university courses.

I was curious if you or your office might be able to assist me with respect to recruitment. I would like to target STEM students either by listserv or through selection in the Registrar's office. So far, recruitment seems to be the most challenging aspect of my study. As such, I would greatly appreciate any help you could provide. Of course, I would be glad to share any data collected from <institution> with you and your office.

If you have questions, please do not hesitate to ask.

Thanks,
David Little

APPENDIX F
Study Invitation Email

Dear Post-secondary Students:

You are being invited to take part in a research study about your engagement in various issues concerning sustainability. You were selected because you are majoring in a science, technology, engineering, mathematics, or social science field. These majors play a crucial role in sustainability related issues.

Although you will not get personal benefit from taking part in this research study, your responses may help educators and researchers work towards building new knowledge for sustainability and sustainability education.

We hope to receive completed questionnaires from about 1,500 people, so your answers are important to us. Of course, you have a choice about whether or not to complete the survey, but if you do participate, you are free to skip any questions or discontinue at any time.

The survey can be found by clicking the link below. The survey will take about 5 to 10 minutes to complete.
<insert link>

There are no known risks to participating in this study.

You will not be asked for any personally identifying information. Your response to the survey will be kept confidential to the extent allowed by law. When we write about the study you will not be identified.

Please be aware, while we make every effort to safeguard your data once received from the online survey/data gathering company, given the nature of online surveys, as with anything involving the Internet, we can never guarantee the confidentiality of the data while still on the survey/data gathering company's servers, or while en route to either them or us.

If you have questions, suggestions, concerns, or complaints about the study, you can contact David Little via email at david.little@uky.edu or via phone at 859-536-2044. If you have any questions about your rights as a volunteer in this research, contact the staff in the Office of Research Integrity at the University of Kentucky between the business hours of 8am and 5pm EST, Mon-Fri. at 859-257-9428 or toll free at 1-866-400-9428.

Thank you in advance for your assistance with this important project. To ensure your responses will be included, please complete the survey by 11:59pm on Feb. 28th, 2014.

Sincerely,
David L. Little II
Science, Technology, Engineering, and Mathematics Education

College of Education
University of Kentucky
PHONE: 859-536-2044
E-MAIL: david.little@uky.edu

APPENDIX G
Phase 3: Student Interview Protocol

Instructions (read to participant)

Thank you for participating in this study. I am going to ask you some questions concerning how you think about sustainability. Please ask if you have questions at any time. You are free to discontinue at any point if you wish. Do you have any questions before we begin?

Questions

1. When you hear the word "sustainability", what comes to mind?
2. How do you think the environment ties in to sustainability?
3. How do you think society ties in to sustainability?
4. How do you think economics ties in to sustainability?
5. Do you think there are any differences in how any of these three (environment, society, and economics) tie in to sustainability? If so, can you describe those differences?
    a. Why do you think <the environment, society, economics> are different from the other two?
    b. Do you think your major prepares you to address the needs of <the environment, society, economics> when it comes to sustainability? Why or why not?
    c. Would you want to work in a position that addresses the needs of <the environment, society, economics> when it comes to sustainability? Why or why not?

APPENDIX H
Phase 3, Round 1: Initial STEMSEI Instrument

Instructions: Below are a series of statements about sustainability. For each statement, please choose one option that best reflects your level of **concern** regarding the statement. There are no right or wrong answers.

Items:
How much concern do you have about:

1. Businesses improving product development?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
2. Businesses being efficient in product manufacturing?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
3. Businesses lowering production costs?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
4. Finding renewable resources?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
5. The rate at which humans consume non-renewable resources?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
6. Pollution of the environment?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
7. Educating people globally on how to meet their own basic food needs?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern

8. Offering affordable healthcare access to all?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern
9. Global population growth?
    - No concern
    - Small concern
    - Moderate concern
    - Big concern

Instructions: Below are a series of statements about sustainability. For each statement, please choose one option that best reflects your level of **agreement** with the statement. There are no right or wrong answers.

Items:

10. I am interested in using my major to help businesses operate efficiently.
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree
11. I get satisfaction out of using my major to solve environmental issues.
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree
12. I feel excited when I use my major to solve social problems.
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree
13. I find it interesting to know what my own carbon footprint is.
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree
14. I am not worried about finding renewable sources of electricity.
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree

15. It is too hard to find ways I can contribute to the well-being of people in other countries.
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree

Instructions: Below are a series of statements about sustainability. For each statement, please choose one option that best reflects **how much you know** about the following topics. There are no right or wrong answers.

Items:

How much do you know about:

1. Ways to lowering production costs for businesses?
    - Nothing at all
    - A little
    - A fair amount
    - A lot

2. Making production manufacturing more efficient for businesses?
    - Nothing at all
    - A little
    - A fair amount
    - A lot

3. Improving product development for businesses?
    - Nothing at all
    - A little
    - A fair amount
    - A lot

4. Current renewable energy programs?
    - Nothing at all
    - A little
    - A fair amount
    - A lot

5. Current consumption rates of fossil fuels?
    - Nothing at all
    - A little
    - A fair amount
    - A lot

6. The effects of pollution on the environment?
    - Nothing at all
    - A little
    - A fair amount
    - A lot

7. The effects of global population growth?
   - Nothing at all
   - A little
   - A fair amount
   - A lot
8. Ways to offering affordable healthcare to all?
   - Nothing at all
   - A little
   - A fair amount
   - A lot

9. Global education programs for developing countries?
   - Nothing at all
   - A little
   - A fair amount
   - A lot

Instructions: Below are a series of statements about sustainability. For each statement, please choose one option that best reflects **the extent you think your major could develop solutions** for the following topics. There are no right or wrong answers.

Items:

To what extent do you think your major could help you find solutions for:

10. Helping businesses develop desirable products?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent
11. Making the production of electronics more efficient?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

12. Helping a business stay profitable?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

13. Maximizing the efficient use of renewable resources?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

14. Analyzing the consumption rate of non-renewable resources?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent
15. Measuring pollution emissions?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

16. Finding affordable ways to offer quality healthcare?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

17. Educating others about their carbon footprint?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

18. Finding safe ways to mass produce food?
    - To no extent
    - To a small extent
    - To some extent
    - To a great extent

Demographic Items - *Options in brackets where applicable.*

1) Are you an undergraduate student or a graduate student? <radio dial> {Undergraduate, Graduate}

2) In what year do you currently plan to graduate? <radio dial> {Already graduated, 2014, 2015, 2016, 2017, 2018, 2019 or beyond}

3) Please select the your gender: <radio dial> {female, male}

4) Please indicate your major: <radio dial> {list composed of all majors listed on the University of Kentucky STEM Majors sheet, which is attached}

5) Please indicate your age: <text box>

6) Please indicate your race: <radio dial> {American Indian/Alaska Native, Asian, Black/African American, Hispanic, Native Hawaiian/Pacific Islander, White/Caucasian, Other [specify]}

7) Please indicate your ethnicity: <radio dial selection> {Hispanic/Latino, Non-Hispanic/non-Latino}

8) Please indicate your country of origin: <text box>

9) The residence in which I grew up practiced the following religion(s): <text box> [Religious affiliation and spiritual beliefs have been posited to impact sustainability perceptions (Lalonde & Jackson, 2002; Thapa, 1999).]

10) I currently believe in or practice the following religion(s): <text box> [Religious affiliation and spiritual beliefs have been posited to impact sustainability perceptions (Lalonde & Jackson, 2002; Thapa, 1999).]

11) Please indicate your sexual orientation: <radio dial> {Heterosexual, Homosexual, Bisexual, Asexual, Prefer not to respond}

APPENDIX I
Phase 3, Round 2: Final STEMSEI Instrument

Instructions: Below are a series of statements about sustainability. For each statement, please choose one option that best reflects your level of **concern/agreement/knowledge** regarding the statement. There are no right or wrong answers.

Final STEMSEI Instrument - Economic Items

Emotional Subscale	1	How much concern do you have about businesses lowering production costs?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	2	How much concern do you have about businesses being efficient in product manufacturing?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	3	How much concern do you have about businesses improving product development?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	4	How much concern do you have about businesses staying profitable?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	5	How much concern do you have about the United States economy being stable?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	6	How much concern do you have about the United States economy growing?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	7	I am interested in using my major to help businesses operate efficiently.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
Cognitive (Knowing)	8	How much do you know about improving product development for businesses?			
		Nothing at All	A Little	A Fair Amount	A Lot
	9	How much do you know about making production manufacturing more efficient for businesses?			
		Nothing at All	A Little	A Fair Amount	A Lot
	10	How much do you know about ways to lower production costs for businesses?			
		Nothing at All	A Little	A Fair Amount	A Lot
	11	How much do you know about how businesses stay profitable?			
		Nothing at All	A Little	A Fair Amount	A Lot

Cognitive (Knowing) Subscale (cont.)	12	How much do you know about how the stability of the United States economy is assessed?			
		Nothing at All	A Little	A Fair Amount	A Lot
	13	How much do you know about how the growth of the United States economy is assessed?			
		Nothing at All	A Little	A Fair Amount	A Lot
	14	How much do you know about how the stability of economies in other countries are assessed?			
		Nothing at All	A Little	A Fair Amount	A Lot
Cognitive (Solving) Subscale	15	To what extent could your major help businesses develop desirable products?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	16	To what extent could your major help businesses stay profitable?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	17	To what extent could your major help make the production of electronics more efficient?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	18	To what extent could your major analyze trends in the United States economy?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	19	To what extent could your major help businesses find new ways to grow?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	20	To what extent could your major help analyze trends in global economies?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	21	To what extent could your major help businesses analyze cost-return ratios?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
Omitted	1	How much concern do you have about the stability of economies in other countries?			
		No Concern	Little Concern	Moderate Concern	Big Concern

Final STEMSEI Instrument - Environmental Items					
Emotional Subscale	1	How much concern do you have about pollution of the environment?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	2	How much concern do you have about finding renewable resources?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	3	How much concern do you have about understanding if climate change is happening?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	4	I enjoy using my major to solve environmental issues.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
	5	I find it interesting to know what my own carbon footprint is.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
	6	I would be comfortable having a limited amount of water I could use daily.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
Cognitive (Knowing) Subscale	7	How much do you know about current consumption rates of fossil fuels?			
		Nothing at All	A Little	A Fair Amount	A Lot
	8	How much do you know about current renewable energy programs?			
		Nothing at All	A Little	A Fair Amount	A Lot
	9	How much do you know about the ecological systems where you currently live?			
		Nothing at All	A Little	A Fair Amount	A Lot
	10	How much do you know about how fossil fuels are refined?			
		Nothing at All	A Little	A Fair Amount	A Lot
	11	How much do you know about how fossil fuels are extracted from the earth?			
		Nothing at All	A Little	A Fair Amount	A Lot
	12	How much do you know about how you use electricity in your home?			
		Nothing at All	A Little	A Fair Amount	A Lot
	13	How much do you know about how electricity is produced?			
		Nothing at All	A Little	A Fair Amount	A Lot

Cognitive (Solving) Subscale	14	To what extent could your major help analyze the consumption rate of non-renewable resources?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	15	To what extent could your major help maximize the efficient use of renewable resources?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	16	To what extent could your major help measure pollution emissions?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	17	To what extent could your major help make the use of fossil fuels more efficient?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	18	To what extent could your major help measure environmental effects from electricity production?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	19	To what extent could your major help maximize efficient use of electricity use in homes?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
Omitted	1	How much concern do you have about the rate at which humans consume non-renewable resources?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	2	I would be comfortable having a limited amount of electricity I could use daily.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
	3	How much do you know about the effects of pollution on the environment?			
		Nothing at All	A Little	A Fair Amount	A Lot
	4	I worry about sources of fossil fuels being depleted in my lifetime.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
	5	How much do you know about how humans increase or decrease biodiversity in the area where you currently live?			
		Nothing at All	A Little	A Fair Amount	A Lot
	6	To what extent could your major help minimize decreases to biodiversity?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	7	To what extent could your major help analyze how ecological systems respond to various changes?			

		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
Final STEMSEI Instrument - Social Items					
Emotional Subscale	1	How much concern do you have about affordable healthcare options being available to you?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	2	How much concern do you have about educating people globally on how to meet their own basic food needs?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	3	How much concern do you have about offering affordable healthcare access to all?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	4	How much concern do you have about the population growth in other countries?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	5	I get satisfaction from contributing to the well-being of people in other countries.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
	6	It is important to use my major to solve social problems.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
Cognitive (Knowing) Subscale	7	How much do you know about the effects of global population growth?			
		Nothing at All	A Little	A Fair Amount	A Lot
	8	How much do you know about global education programs for developing countries?			
		Nothing at All	A Little	A Fair Amount	A Lot
	9	How much do you know about ways to offering affordable healthcare to all?			
		Nothing at All	A Little	A Fair Amount	A Lot
	10	How much do you know about the social needs of diverse groups in the United States?			
		Nothing at All	A Little	A Fair Amount	A Lot
	11	How much do you know about the social needs of diverse groups in other countries?			
		Nothing at All	A Little	A Fair Amount	A Lot
	12	How much do you know about the education efforts for sustainability in the United States?			

		Nothing at All	A Little	A Fair Amount	A Lot
	13	How much do you know about the education efforts for sustainability in other countries?			
		Nothing at All	A Little	A Fair Amount	A Lot
	14	How much do you know about the healthcare options available to people in third-world countries?			
		Nothing at All	A Little	A Fair Amount	A Lot
Cognitive (Solving) Subscale	15	To what extent could your major help develop solutions for safely mass producing food?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	16	To what extent could your major help develop affordable ways to offer quality healthcare?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	17	To what extent could your major improve how we provide medical care in the United States?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	18	To what extent could your major improve how medical care is provided in third-world countries?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	19	To what extent could your major improve how we address diverse social needs in the United States?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
Omitted	1	How much concern do you have about the population growth in the United States?			
		No Concern	Little Concern	Moderate Concern	Big Concern
	2	I get satisfaction from contributing to the well-being of people in my neighborhood.			
		Strongly Disagree	Disagree	Agree	Strongly Agree
	3	How much do you know about the population growth rate in the United States?			
		Nothing at All	A Little	A Fair Amount	A Lot
	4	To what extent could your major improve how diverse social needs are addressed in other countries?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent
	5	To what extent could your major help educate others about their carbon footprint?			
		To No Extent	To a Small Extent	To Some Extent	To a Great Extent

6	To what extent could your major measure various effects from the global population rate?			
	To No Extent	To a Small Extent	To Some Extent	To a Great Extent

Final STEMSEI Instrument - Demographic Items

1) Are you an undergraduate student or a graduate student? <radio dial> {Undergraduate, Graduate}

2) In what year do you currently plan to graduate? <radio dial> {Already graduated, 2014, 2015, 2016, 2017, 2018, 2019 or beyond}

3) Please select the your gender: <radio dial> {female, male}

4) Please indicate your major: <radio dial> {list of all possible STEM majors at corresponding institution}

5) Please indicate your age: <text box>

6) Please indicate your race: <radio dial> {American Indian/Alaska Native, Asian, Black/African American, Hispanic, Native Hawaiian/Pacific Islander, White/Caucasian, Mixed [specify], Other [specify]}

7) Please indicate your ethnicity: <radio dial selection> {Hispanic/Latino, Non-Hispanic/non-Latino}

8) Please indicate your country of origin: <text box>

9) The residence in which I grew up practiced the following religion(s): <text box>

10) I currently believe in or practice the following religion(s): <text box>

11) Please indicate your sexual orientation: <radio dial> {Heterosexual, Homosexual, Bisexual, Asexual, Prefer not to respond}

REFERENCES

Aighewi, I. T., & Osaigbovo, U. A. (2009). Students' perspectives on worldwide "greening" of tertiary education curricula. *Research in Science Education, 40*(5), 625-637. doi:10.1007/s11165-009-9137-x

Alwin, D. F., (2010). How good is survey measurement? Assessing the reliability and validity of survey measures. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 405-434). Wagon Lane, Bingley, United Kingdom: Emerald Group Publishing.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Ameer, R., & Othman, R. (2012). Sustainability practices and corporate financial performance: A study based on the top global corporations. *Journal of Business Ethics, 108*(1), 61-79. doi:10.1007/s10551-011-1063-y

Anderson, R. (2009, May). *Ray Anderson on the business logic of sustainability*. [Video]. Retrieved from http://www.ted.com/talks/lang/en/ray_anderson_on_the_business_logic_of_sustainability.html

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology, 44* , 427–445. doi: 10.1016/j.jsp. 2006.04.002 .

Arrow, K. J., Dasgupta, P., Goulder, L. H., Mumford, K. J., & Oleson, K. (2012). Sustainability and the measurement of wealth. *Environment and Development Economics, 17*, 317-353.

Astin, A. W., & Antonio, A. L. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education.* Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

Attenborough, D. (2011). This heaving planet. *New Statesman, 140*, 28-32.

Badurdeen, F., Brown, A., Gregory, R., Fu, H., Schroeder, M., Sekulic, D., Vincent, L. & Luhan, G. (2013). Reframing interdisciplinary approaches to systems thinking for sustainability. In M. Bilec & J. Chol (Eds.), *Proceedings of the International Symposium on Sustainable Systems & Technologies* (ISSN 2329-9169).

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, Maryland: Eric Clearing House on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED 458 219).

Baker, R. P., Crawford, S., & Swinehart, J. (2004). Development and testing of web questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E.

Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 361-384). Hoboken, New Jersey: John Wiley & Sons, Inc.

Bandalos, D., & Kopp, J. (2013, May). *The utility of exploratory bi-factor rotations in scale construction.* Paper presented at the annual meeting of the American Psychological Society, Washington, D.C.

Basiago, A. D. (1995). Methods of defining "sustainability". *Sustainable Development, 3,* 109-119.

Bell, S., & Morse, S. (2008). *Sustainability indicators: Measuring the immeasurable?* [Adobe Digital Editions version]. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=225261&site=ehost-live&scope=site

Bencze, L., & Carter, L. (2011). Globalizing students acting for the common good. *Journal of Research in Science Teaching, 48*(6), 648-669. doi:10.1002/tea.20419

Bencze, L., Sperling, E., & Carter, L. (2011). Students' research-informed socio-scientific activism: Re/visions for a sustainable future. *Research in Science Education*, *42*, 1-20. doi:10.1007/s11165-011-9260-3

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *The American Journal of Occupational Therapy, 36,* 789-800.

Bohrnstedt, G. W. (2010). Measurement models for survey research. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 347-404). Wagon Lane, Bingley, United Kingdom: Emerald Group Publishing.

Borghesi, S., & Vercelli, A. (2008). *Global sustainability: Social and environmental conditions.* [Adobe Digital Editions version]. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=266596&site=ehost-live&scope=site

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York, New York: Springer. Retrieved from http://www.mun.ca/biology/quant/ModelSelectionMultimodelInference.pdf

British Broadcasting Corporation. (2004). *BBC - Religions - Unitarianism: Unitarianism at a glance.* Retrieved from http://www.bbc.co.uk/religion/religions/unitarianism/ataglance/glance.shtml

Brundiers, K., & Wiek, A. (2011). Educating students in real-world sustainability research: Vision and implementation. *Innovative Higher Education, 36*(2), 107-124. doi:10.1007/s10755-010-9161-9

Brown, A. S. (2011). ASME's third annual survey finds that engineers are still trying to understand how sustainability fits into their workflow. *Mechanical Engineering, 133*(11), 36-41.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: User guide*. Lincolnwood, IL: Scientific Software International, Inc.

Carley, K. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology, 23*, 75-126. Retrieved from http://www.casos.cs.cmu.edu/publications/papers/carley_1993_codingchoices.PDF

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Christensen, P., Thrane, M., Jorgensen, T. H., & Lehmann, M. (2009). Sustainable development: Assessing the gap between preaching and practice at Aalborg University. *International Journal of Sustainability in Higher Education, 10*(1), 4-20.

Conrad, F., & Blair, J. (1996). *From Impressions to data: Increasing the objectivity of cognitive interviews*. Paper presented at the meeting of the ASA Section on Survey Research Methods, Alexandria, Virginia. Retrieved from http://www.bls.gov/ore/pdf/st960080.pdf

Cordano, M., Welcomer, S. A., & Scherer, R. F. (2003). An analysis of the predictive validity of the New Ecological Paradigm scale. *Journal of Environmental Education, 34*(3), 22-28.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, New Jersey: Merrill Prentice Hall.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 209-240). Thousand Oaks, California: Sage Publications, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/bf02310555

Davis, S. A., Edmister, J. H., Sullivan, K., & West, C. K. (2003). Educating sustainable societies for the twenty-first century. *International Journal of Sustainability in Higher Education, 4*(2), 169-179.

de Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*(2), 155-170.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, New York: Guildford Press.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 361-384). Hoboken, New Jersey: John Wiley & Sons, Inc.

DeMars, C. E. (2013). A tutoring on interpreting bifactor model scores. *International Journal of Testing, 13,* 354-378. doi: 10.1080/15305058.2013.799067

DeVellis, R. F. (2003). *Scale development: Theory and applications.* Thousand Oaks, California: Sage Publications, Inc. Retrieved from http://www.bus.ucf.edu/faculty/rfolger/file.axd?file=2012%2F11%2FDeVellis+-+Scale+development.pdf

Djordjevic, A., & Cotton, D. R. E. (2011). Communicating the sustainability message in higher education institutions. *International Journal of Sustainability in Higher Education, 12*(4), 381-394. doi:10.1108/14676371111168296

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3-26). Hillsdale, New Jersey: Lawrence Erlbaum.

Duchin, F. (1996). Population change, lifestyle, and technology: How much difference can they make? *Population Council, 22*(2), 321-330. doi: 10.2307/2137437.

Dunlap, R. E. (2008). The new environmental paradigm scale: From marginality to worldwide use. *Journal of Environmental Education, 40*(1), 3-18.

Dunlap, R. & Van Liere, K. (1978). The “new environmental paradigm”: A proposed measuring instrument and preliminary results. *The Journal of Environmental Education, 9*(4), 10–19.

Dunlap, R. E., Van Liere, K. D., Mertig, A. G., & Jones, R. E. (2000). Measuring endorsement of the new ecological paradigm: A revised NEP scale. *Journal of Social Issues, 56*(3), 425-442.

Dyllick, T., & Hockerts, K. (2002). Beyond the business case for corporate sustainability. *Business Strategy & the Environment (John Wiley & Sons, Inc), 11*(2), 130-141. doi:10.1002/bse.323

Erdogan, M. (2010). Sustainability in higher education: A needs assessment on a course "education and awareness for sustainability". *Asia-Pacific Forum on Science Learning and Teaching, 11*(1).

Erzberger, C., & Kelle, U. (2003). Making inferences in mixed methods: The rules of integration. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 457-488). Thousand Oaks, California: Sage Publications, Inc.

Fairbrass, J., & Zueva-Owens, A. (2011). Conceptualising corporate social responsibility: 'Relational governance' assessed, augmented, and adapted. *Journal of Business Ethics, 105*, 321-335. doi:10.1007/s10551-011-0968-9

Fan, X., Miller, B. C., Park, K., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods, 18*, 223-244.

Franck, S., von Bloh, W., Müller, C., Bondeau, A., & Sakschewski, B. (2011). Harvesting the sun: New estimations of the maximum population of planet earth. *Ecological Modeling, 222*(12), 2019-2026. doi:10.1016/j.ecolmodel.2011.03.030

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Frisk, E., & Larson, K. L. (2011). Educating for sustainability: Competencies & practices for transformative action. *Journal of Sustainability Education, 2.* Retrieved from http://www.jsedimensions.org/wordpress/wp-content/uploads/2011/03/FriskLarson2011.pdf

Gagnon, B., Leduc, R., & Savard, L. (2009). Sustainable development in engineering: A review of principles and definition of a conceptual framework. *Environmental Engineering Science, 26*(10), 1459-1472. doi:10.1089/ees.2008.0345

Gibbon, D., Lake, A., & Stocking, M. (1996). Sustainable development: A challenge for agriculture. In S. Morse & M. Stocking (Eds.), *People and environment* (pp. 31-68): Vancouver, British Columbia, Canada: University of British Columbia Press.

Gonzalez, H. B., & Kuenzi, J. J. (2012). *Science, technology, engineering, and mathematics (STEM) education: A primer*. Retrieved from Federation of American Scientists website: http://www.fas.org/sgp/crs/misc/R42642.pdf.

Griffin, K. A., Perez II, D., Holmes, A. P. E., & Mayo, C. E. P. (2010). Investing in the future: The importance of faculty mentoring in the development of students of color in STEM. *New Directions for Institutional Research, 2010*(148), 95-103. doi: 10.1002/ir.365

Hanson, S. L. (2004). African American women in science: Experiences from high school through the post-secondary years and beyond. *National Women's Studies Association Journal, 16*(1), 96-115.

Hediger, W. (2006). Weak and strong sustainability, environmental conservation and economic growth. *Natural Resource Modeling, 19*(3), 359-394. doi:http://www3.interscience.wiley.com/journal/118902516/home

Hodson, D. (2003). Time for action: science education for an alternative future. *International Journal of Science Education*, 25(6), 645–670.

Holgado–Tello, F., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. [Article]. *Quality & Quantity, 44*(1), 153-166. doi: 10.1007/s11135-008-9190-y

Hrebicek, J., Valtinyova, S., Kren, J., Hodinka, M., Trenz, O., & Marada, P. (2013). Sustainability indicators: Development and application for the agriculture sector. In M. G. Erechtchoukova, P. A. Khaiter, & P. Golinska (Eds.), *Sustainability appraisal: Quantitative methods and mathematical techniques for environmental performance evaluation* (pp. 63-103). New York, NewYork: Springer.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118

Hughes, C. S., & Estes, C. A. (2005). The influence of environmental education on environmentally responsible behaviors of undergraduate students in a traditional and nontraditional setting. *Journal of Experiential Education, 27*(3), 308-310.

IBM Corp. (2011). IBM SPSS Statistics for Windows (Version 20.0) [Software]. Armonk, New York: IBM Corp.

Izac, A. M. N., & Swift, M. J. (1994). On agricultural sustainability and its measurement in small-scale farming in sub-Saharan Africa. *Ecological Economics, 11*(2), 105-125.

Jennrich, R. I., & Bentler, P. M. (2011a). Exploratory bi-factor analysis. *Psychometrika, 76*, 537-549. doi: 10.1007/s11336-011-9218-4

Jennrich, R. I., & Bentler, P. M. (2011b). *Exploratory bi-factor analysis: The oblique case.* Unpublished manuscript, Department of Statistics, University of California, Los Angeles, Los Angeles, California. Retrieved from http://statistics.ucla.edu/system/resources/BAhbBlsHOgZmSSI~MjAxMi8wNS8xNi8xNF8xOF8zOF82OTBfRXhwbG9yYXRvcnlfQmlfZmFjdG9yX0FuYWx5c2lzLnBkZgY6BkVU/Exploratory%20Bi-factor%20Analysis.pdf

Johnson, C. Y., Bowker, J. M., & Cordell, H. K. (2004). Ethnic variation in environmental belief and behavior: An examination of the new ecological paradigm in a social psychological context. *Environment and Behavior, 36*(2), 157-186.

Johnston, P., Everard, M., Santillo, D., & Robèrt, K. H. (2007). Reclaiming the definition of sustainability. *Environmental Science and Pollution Research International, 14*(1), 60-66.

Kararni, H. (2012). An introduction to differential item functioning. *The International Joural of Educational and Psychological Assessment, 11*(2), 59-76.

Kidd, C. V. (1992). The evolution of sustainability. *Journal of Agricultural & Environmental Ethics, 5*(1), 1-26.

Kiewiet, D. J., & Vos, J. F. J. (2007). Organisational sustainability: A case for formulating a tailor-made definition. *Journal of Environmental Assessment Policy & Management, 9*(1), 1-18.

Kline, P. (2000). *Handbook of psychological testing.* New York, New York: Routledge.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263-313). Wagon Lane, Bingley, United Kingdom: Emerald Group Publishing.

Lalonde, R., & Jackson, E. L. (2002). The New Environmental Paradigm scale: Has it outlived its usefulness? *Journal of Environmental Education, 33*(4), 28-36.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation). Retrieved from ProQuest.

Le Grange, L. L. L. (2011). Sustainability and higher education: From arborescent to rhizomatic thinking. *Educational Philosophy & Theory, 43*(7), 742-754. doi:10.1111/j.1469-5812.2008.00503.x

Lemonick, M. D. (2009). Top 10 myths about sustainability. *Scientific American Earth 3.0, 19*(1), 40-45.

Linn, R. L. (2006). *The standards for educational and psychological testing: Guidance in test development.* In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3-26). Hillsdale, New Jersey: Lawrence Erlbaum

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum.

Mann, S., Harraway, J., Broughton-Ansin, F., Deaker, L., & Shephard, K. (2013). Seeking richer descriptions of learner's sustainability attributes and learning needs. *International Journal of Sustainability in Higher Education, 14*(1), 90-100.

Matteuccia, M., Mignania, S., & Veldkamp, B. P. (2012). Prior distributions for item parameters in IRT models. *Communications in Statistics: Theory and Methods, 41*, 2944-2958. doi:10.1080/03610926.2011.639973

Marshall, J. D., & Toffel, M. W. (2004). Framing the elusive concept of sustainability: A sustainability hierarchy. *Environmental Science & Technology, 39*(3), 673-682. doi:10.1021/es040394k

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research, 3*(1), 97-110.

McDonough, W., & Braungart, M. (2002). *Cradle to cradle : Remaking the way we make things*. New York, New York: North Point Press.

McMillan, J. H., & Schumacher, S. (2010). *Research in education: Evidence-based inquiry.* Upper Saddle River, New Jersey: Pearson Education, Inc.

Meadows, D. H. (2008). *Thinking in systems: a primer.* White River Junction, Vermont: Chelsea Green Publishing.

Mertens, D. M. (2005). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods.* Thousand Oaks, California: Sage Publications, Inc.

Messick, S. (1990). *Validity of test interpretation and use* (ETS Research Report 90-11). Retrieved from http://files.eric.ed.gov/fulltext/ED395031.pdf.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation.* Thousand Oaks, California: Sage Publications, Inc.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2009). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York, New York: Guilford Press.

Morris, M. (2012). Sustainability: An exercise in futility. *International Journal of Business & Management, 7*(2), 36-44. doi:10.5539/ijbm.v7n2p36

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2), 1-19. Retrieved from http://www.ualberta.ca/~iiqm/backissues/1_2Final/pdf/morseetal.pdf

Munier, N. (2005). *Introduction to sustainability: Road to a better future*. [Adobe Digital Editions version]. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=143378&site=ehost-live&scope=site

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Muthén, L. K., & Muthén, B. O. (1998-2010). Mplus user's guide (6th ed.). Los Angeles, CA: Authors.

National Research Council. (1999). *Our common journey : A transition toward sustainability*. Washington, D.C.: National Academy Press.

National Science Foundation. (2006). *Graduate research fellowship program (GRFP).* Retrieved from http://www.nsf.gov/pubs/2012/nsf12599/nsf12599.htm#appendix.

National Science Board. (2010). *Preparing the next generate of STEM innovators: Identifying and developing our nation's human capital* (NSB Publication No. NSB-10-33). Retrieved from https://www.nsf.gov/nsb/publications/2010/nsb1033.pdf

Nattrass, B. F., & Altomare, M. (1999). The natural step for business wealth, ecology and the evolutionary corporation. [Adobe Digital Editions version]. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=478316

Norcini, J. J., Day, S. C., Grosso, L. J., Langdon, L. O., Kimball, H. R., Popp, R. L., & Goldfinger, S. E. (1993). The relevance to clinical practice of the certifying examination in internal medicine. *Journal of General Internal Medicine, 8*(2), 82-85.

Norman, W., & MacDonald, C. (2004). Getting to the bottom of "triple bottom line". *Business Ethics Quarterly, 14*(2), 243-262.

O'Connell, T. S., Potter, T. G., Curthoys, L. P., Dyment, J. E., & Cuthbertson, B. (2005). A call for sustainability education in post-secondary outdoor recreation programs. *International Journal of Sustainability in Higher Education, 6*, 81-94.

Olli, E., Grendstad, G., & Wollebaek, D. (2001). Correlates of environmental behaviors: Bringing back social context. *Environment and Behavior, 33*(2), 181-208.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Okoli, C., & Pawlowski, S. D. (2003). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management, 42*(1), 1-20. Retrieved from

http://spectrum.library.concordia.ca/976864/1/OkoliPawlowski2004DelphiPostprint.pdf

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24, 50-64. doi:10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-□2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement, 27, 289-298. doi:10.1177/0146621603027004004

Palmer, R. T., Maramba, D. C., Elon Dancy, T. (2011). A qualitative investigation of factors promoting the retention and persistence of students of color in STEM. *Journal of Negro Education, 80*(4), 491-504.

Pezzoli, K. (1997). Sustainable development: A transdisciplinary view of the literature. *Journal of Environmental Planning and Development, 40*, 549-574.

Quental, N., Lourenço, J., & da Silva, F. (2011). Sustainability: Characteristics and scientific roots. *Environment, Development & Sustainability, 13*(2), 257-276. doi:10.1007/s10668-010-9260-x

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, New York: Springer. doi: 10.1007/978-0-387-89976-3

Reise, S. P., Moore, T. M., & Haviland, M. G. (2013). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559.

Robinson, J. (2004). Squaring the circle? Some thoughts on the idea of sustainable development. *Ecological Economics, 48*(4), 369-384. doi:10.1016/j.ecolecon.2003.10.017

Ruttan, V. W. (1994). Constraints on the design of sustainable systems of agricultural production. *Ecological Economics, 10*(3), 209-219. doi:10.1016/0921-8009(94)90109-0

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Schaller, N. (1989). Low input sustainable agriculture, in Smith, D.T. (1989). *Yearbook of Agriculture: Farm Management*, US Department of Agriculture, Washington, DC, 216 -219.

Schmid J, & Leiman J. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Schneiderman, D., & Freihoefer, K. (2012). A pre- and post-evaluation of integrating sustainability curriculum by inserting Okala modules into an interior design materials and methods course. *International Journal of Sustainability in Higher Education*, *13*, 408-423.

Sekulic, D. P. (2011). An entropy-based metric for a transformational technology development. In B. R. Bakshi, T. G. Gutowski, & D. P. Sekulic (Eds.), *Thermodynamics and the destruction of resources* (pp. 133-162). New York, New York: Cambridge University Press.

Seliger, G. Khraisheh, M. M. K., & Jawahir, I. S. (2011). *Advances in sustainable manufacturing.* New York, New York: Springer.

Shriberg, M. (2002). Institutional assessment tools for sustainability in higher education: Strengths, weaknesses, and implications for practice and theory, *Higher Education Policy*, *15*, 153-167.

Smeding, A. (2012). Women in science, technology, engineering, and mathematics (STEM): An investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance. *Sex Roles*, *67*, 617-629.

Smith, E. T. (2004a). *Water resources sustainability*. Carbondale, Illinois: Universities Council on Water Resources. Retrieved from http://www.awra.org/committees/techcom/watershed/pdfs/AWRA%20H&WM_2010_1_what-is-sustainability.pdf

Smith, P. A. C., & Sharicz, C. (2011). The shift needed for sustainability. *Learning Organization*, *18*(1), 73-86.

Smith, T. W., (2004b). Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 431-452). Hoboken, New Jersey: John Wiley & Sons, Inc.

Solow, R. M. (1974). Intergenerational equity and exhaustible resources. *The Review of Economic Studies*, *41*, 29-45.

Solow, R. M. (1986). On the intergenerational allocation of natural resources. *The Scandinavian Journal of Economics*, *88*, 141-149.

Stark, S. (2008). MODFIT (Version 3.0) [Software]. Urbana-Champaign: University of Illinois, Urbana-Champaign.

Talbot, J., & Venkataraman, R. (2011). Integration of sustainability principles into project baselines using a comprehensive indicator set. *International Business & Economics Research Journal*, *10*(9), 29-40.

Thapa, B. (1999). Environmentalism: The relation of environmental attitudes and environmentally responsible behaviors among undergraduate students. *Bulletin of Science, Technology & Society, 19*(5), 426.

Trzyna, T. C. (1995). *A sustainable world: Defining and measuring sustainable development.* Sacramento, California: International Center for the Environment and Public Policy.

United Nations Division for Sustainable Development. (2001). *Indicators for sustainable development: Framework and methodologies.* Retrieved from http://www.un.org/esa/sustdev/csd/csd9_indi_bp3.pdf.

United Nations Economic Commission for Europe. (2005). *UNECE annual report 2004-2005: Sustainable development – concept and action.* Retrieved from http://www.unece.org/oes/nutshell/2004-2005/focus_sustainable_development.html

United Nations Educational, Scientific and Cultural Organization. (2005, October). *United Nations decade of education for sustainable development (2005-2014): International implementation scheme*. Retrieved from http://unesdoc.unesco.org/images/0014/001486/148654e.pdf

United States Department of Agriculture. (n. d.) *NSRE Survey Questions.* Retrieved from http://www.srs.fs.usda.gov/trends/nsre-directory/questions.html.

Van Cauwenbergh, N., Biala, K., Bielders, C., Brouckaert, V., Franchois, L., Garcia Cidad, V., . . . Peeters, A. (2007). Safe—a hierarchical framework for assessing the sustainability of agricultural systems. *Agriculture, Ecosystems & Environment, 120*(2), 229-242. doi:10.1016/j.agee.2006.09.006

Van Kerkhoff, L., & Lebel, L. (2006). Linking knowledge and action for sustainable development. *Annual Review of Environment & Resources, 31*(1), 445-477. doi:10.1146/annurev.energy.31.102405.170850

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, Massachusetts: Harvard University Press.

Wæhrens, E. E. (2010). *Measuring quality of occupational performance based on self-report and observation: Development and validation of instruments to evaluate ADL task performance.* Retrieved from http://www.diva-portal.org/smash/get/diva2:318271/FULLTEXT01.pdf

Wainwright, J. (2010). Climate change, capitalism, and the challenge of transdisciplinarity. *Annals of the Association of American Geographers, 100*, 983-991. doi:10.1080/00045608.2010.502439

Webb, N. L. (1997). *Alignment of math and science standards and assessments in four states* (Research Monograph No. 18). Washington, District of Columbia: Council

of Chief State School Officers. Retrieved from http://facstaff.wceruw.org/normw/WEBBMonograph6criteria.pdf

White, M. A. (2013). Sustainability: I know it when I see it. *Ecological Economics, 86*, 213-217. doi: 10.1016/j.ecolecon.2012.12.020

Willis, G. B. (1999). *Cognitive interviewing: A "How To" guide*. Paper presented at the 1999 Meeting of the American Statistical Association, Baltimore, Maryland.

World Commission on Environment and Development. (1987). *Our common future*. New York, New York: Oxford University Press. Retrieved from http://www.un-documents.net/our-common-future.pdf

Wright, H. A., Ironside, J. E., & Gwynn-Jones, D. (2009). Sustainability in bioscience fieldwork: Practical information from a UK agricultural research institute. *International Journal of Sustainability in Higher Education, 10*(1), 54-67.

Yu, C-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Dissertation. Los Angeles, CA. Retrieved from http://statmodel2.com/download/Yudissertation.pdf

VITA
David L. Little II
Place of Birth: Jackson, Kentucky

Academic Background

University of Kentucky Lexington, Kentucky	MAEd in Secondary Education, Mathematics MA in Mathematics	August 2010 May 2007
Morehead State University Morehead, Kentucky	BS with an Area of Concentration in Mathematics	May 2002

Professional Work Experience

2013 - 2014	Mathematics Department University of Kentucky	Part-time Instructor
2012 - 2013	See Blue STEM Camp University of Kentucky	Researcher, External Evaluator, Co-Director of Academics, and Presenter
2011 - 2013	See Blue Mathematics Clinic University of Kentucky	Researcher and External Evaluator
2010 - 2013	Mathematics Department University of Kentucky	Research Assistant for National Science Foundation funded Geometry Assessment for Secondary Teachers (GAST) Grant
2011 - 2012	STEM Education Department University of Kentucky	College Teaching Practicum
2009 - 2012	Mathematics Department University of Kentucky	Teaching Assistant
Summer 2011	STEM Education Department University of Kentucky	STEM Interdisciplinary Curriculum Developer
2008 - 2009	Mathematics Department The Madeira School McLean, Virginia	High School Mathematics Teacher, Course Coordinator for BC Calculus and Geometry, Faculty Mentor for the Math Team
2007 - 2008	National Cathedral School Washington, D.C.	Middle School Mathematics Teacher, After School Program Coordinator
2006-2007	British School of Washington Washington, D.C.	Mathematics Teacher Consultant
2005 - 2007	Mathematics Department University of Kentucky	Teaching Assistant, MathExcel Instructor

2003 - 2005	Mathematics Department Morehead State University	Instructor, and Curriculum/Assessment Coordinator for Developmental Mathematics

Educational Honors and Awards

- Magna Cum Laude Graduation Honors, Morehead State University, Spring 2002
- Thomas Edward Fouch Prize in Mathematics, Morehead State University, Spring 2002
- Outstanding Graduating Student in Mathematics, Morehead State University, Spring 2002
- Outstanding Senior Student in Mathematics, Morehead State University, Spring 2002

Teaching Honors and Awards

- Recognition of Excellence Award, Mathematics: Content Knowledge, Spring 2010, Educational Testing Service
- Nominated for Master Teacher Award, Spring 2009, Madeira School

Scholarship

Refereed Publications

Jackson, C., Schroeder, M., & **Little II, D. L.** (accepted – to appear in 2014). Using informal learning environments to prepare preservice teachers. *Teacher Education & Practice.*

Schroeder, M., Jackson, C., Schroeder, D., Miller, M., Walcott, B., & **Little II, D. L.** (accepted with revisions – to appear in 2014). Developing middle school students' interests in STEM via summer learning experiences: See Blue STEM Camp. *School Science and Mathematics.*

National Refereed Conference Proceedings

Little II, D. L., Toland, M. D., & Morales, A. (2014, April). *First item response theory analysis: A new measure of perceived discrimination for sexual minority Latinas.* Paper presented at the meeting of the American Educational Research Association, Philadelphia, Pennsylvania.

Adumat, S. J., Bouwma-Gearhart, J., **Little II, D. L.**, & Bouwma-Gearhart, A. (2011, April). *Modeling-based curriculum and instruction in the undergraduate classroom: Engagement of students as communities of scientists.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.