Spring 2017

Integrating implementation fidelity and learning improvement to enhance students’ ethical reasoning abilities

Kristen L. Smith
James Madison University

Follow this and additional works at: http://commons.lib.jmu.edu/diss201019

Part of the Curriculum and Instruction Commons, Educational Assessment, Evaluation, and Research Commons, Educational Methods Commons, Higher Education Commons, and the Psychology Commons

Recommended Citation
Smith, Kristen L., "Integrating implementation fidelity and learning improvement to enhance students’ ethical reasoning abilities" (2017). Dissertations. 153.
http://commons.lib.jmu.edu/diss201019/153

This Dissertation is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Dissertations by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
Acknowledgments

I am very grateful to the many people who supported me throughout this dissertation project. First and foremost, I would like to thank my mentor, Dr. Keston Fulcher. You’ve taught me many things that I expected to learn about: best practices in assessment, strategies for implementing learning improvement models, how to be a better writer, etc. But you’ve also taught me so much more: everything from evaluating golden age comic books, to changing the world, to being the kind of leader people actually want to follow. Thank you for letting me be true to myself, but also helping me grow into a better person.

I also benefited from working with an outstanding dissertation committee. Dr. William Hawk and Dr. Lori Pyle, from the Madison Collaborative, taught me how to be innovative and make a difference. Thank you both for all of the opportunities you have given me over the past five years, and for your contributions to planning and executing this dissertation project. Dr. Sara Finney, thank you for teaching me (literally) everything I know about implementation fidelity research. Your contributions undoubtedly made this dissertation project much stronger. Dr. Allison Ames, thank you for all of your exceptional work with the Madison Collaborative. Your methodological skills substantially contributed to this dissertation project.

I would also like to thank my fellow (past and current) graduate students for their support and friendship. Thank you Megan Rodgers Good for introducing me to graduate school and serving as my “adviser.” Your abilities to lead and innovate have encouraged me throughout my time in graduate school and will continue to inspire me in my career. Thank you Rochelle Fisher, Bo Bashkov, Jerusha Gerstner, Dan Jurich, and Becca Marsh.
Runyon for “showing me the ropes,” leading by example, and being my friends. To Thai Ong, Courtney Sanders, Scott Strickman, and Liz Hawk Sanchez, thank you for being unique and incredible friends. I am so grateful for each of you and look forward to being lifelong colleagues.

Finally, I’d like to thank my family members, loved ones, and significant other for their support. Nobody has been more supportive than my mom. Thank you, mom, for showing me that I could go to college when others said that I could not. Your support, throughout these past nine years, has made all the difference for me.
# Table of Contents

Acknowledgments ........................................................................................................... ii
List of Tables ......................................................................................................................... viii
List of Figures ......................................................................................................................... ix
Abstract .................................................................................................................................. x

CHAPTER 1 ......................................................................................................................... 1
Introduction ............................................................................................................................ 1
Defining Key Terms ............................................................................................................... 2
Statement of Problem .......................................................................................................... 17

CHAPTER 2 ......................................................................................................................... 19
Literature Review .................................................................................................................. 19
Purpose of Literature Review ............................................................................................... 19
Historical Contexts of Assessment Practice ....................................................................... 19
  Better assessment ≠ Improved learning ............................................................................. 22
Exploring the Disconnect Between Assessment Practices and Learning Improvement ..... 26
  The Level Problem ............................................................................................................. 26
  Unclear and inconsistent communication ......................................................................... 29
  Lack of easily accessible and detailed learning improvement examples .................... 30
  General culture issues ....................................................................................................... 31
  Lack of focus on program theory and educational interventions ................................... 33
Summary ............................................................................................................................... 38
Need for a Model to Integrate Assessment and Learning Improvement Processes .......... 39
  Examples of national-level learning improvement processes ........................................ 39
  Examples of institutional-level learning improvement processes ................................... 40
The Simple Model in Action ............................................................................................... 43
  Understanding Implementation Fidelity Research and its Importance to Learning Improvement ............................................. 45
  An implementation fidelity example ............................................................................... 46
  Five components of implementation fidelity ................................................................. 49
  Implementation fidelity general data collection methods .............................................. 51
  Importance of implementation fidelity research ......................................................... 52
Brief Analysis of Existing Implementation Fidelity Applications in Academic Contexts ......................................................................................................................... 55
  Existing applications in K-12 contexts .......................................................................... 55
  Examples of fidelity checklists .................................................................................... 57
  Existing applications in higher education contexts .................................................... 61
  Examples of fidelity checklists .................................................................................... 63
Need for More Implementation Fidelity Research

Purpose of Current Project

Introduction to the Madison Collaborative: Ethical Reasoning in Action

Madison Collaborative Learning Interventions and Assessments

It’s Complicated

The Madison Collaborative Interactive (MCI)

MC Assessments

Madison Collaborative’s Impact, So Far

Research Questions

RQ 1: Is the observed variance in ER-WR scores mostly due to differences in students’ ER abilities? Are ER-WR scores reliable?

RQ 2: Did students’ ER skills improve from the beginning to the end of the fall semester? Do students who participated in the current study demonstrate greater ER abilities compared to students who did not?

RQ 3: To what extent did the delivered ER intervention differ from the designed ER intervention?

RQ 4: What insights can implementation fidelity data provide to help stakeholders make more accurate inferences from outcomes assessment results? Are naturally occurring differences in implementation fidelity related to outcomes assessment results?

CHAPTER 3

Method

Faculty-Level Educational Intervention

Co-creating an ER intervention and fidelity checklist

Faculty and Student Participants

Madison Collaborative Assessment Instruments

The ethical reasoning writing rubric (ER-WR)

Reliability and validity evidence

The ethical reasoning identification test (ERIT)

Reliability and validity evidence

Data Collection

Implementation Fidelity Data Preparation

ER-WR Data Preparation: Rater Training and Essay Rating

ERIT Data Preparation

SOS Data Preparation

Data Management and Motivation Filtering Results

Pre-test ER-WR
Pre-test ERIT ................................................................. 104
Pre-test motivation filtering .................................................. 104
Post-test ER-WR .............................................................. 105
Post-test ERIT ............................................................... 105
Post-test informed consent .................................................. 105
Post-test motivation filtering ................................................. 105
“Unrateable” ER-WR essay data ........................................... 106
Pre/post matched sample size for ER-WR and ERIT treated separately .......................... 107

CHAPTER 4 .................................................................. 108

Results .......................................................................... 108

Data Analysis and Results ................................................... 108

RQ 1: Is the observed variance in ER-WR scores mostly due to differences in students’ abilities? Are ER-WR scores reliable? .................................................. 108
Rater Harshness ................................................................ 117
Summary .......................................................................... 117

RQ 2: Did students’ ER skills improve from the beginning to the end of the fall semester? Do students who participated in the current study demonstrate greater ER abilities compared to students who did not? .................................................. 118
ER-WR longitudinal .......................................................... 119
ERIT longitudinal .............................................................. 119
ER-WR cross-sectional ....................................................... 120
ERIT cross-sectional .......................................................... 121
Across faculty comparisons ................................................ 122
ER-WR ........................................................................... 123
ERIT ............................................................................... 124
Results compared to similar study ........................................ 125
Summary .......................................................................... 126

RQ 3: To what extent did the delivered ER intervention differ from the designed ER intervention? .................................................................................. 127
Results aggregated at program component level ................... 131
Results aggregated at specific feature level .......................... 133
Summary .......................................................................... 135

RQ 4: What insights can implementation fidelity data provide to help stakeholders make more accurate inferences from outcomes assessment results? Are naturally occurring differences in implementation fidelity related to outcomes assessment results? .................................................................................. 136
Relationships among ER skills, responsiveness, and quality ........................................ 139
Element A: identify ethical issue in its context ..................................................... 146
Element B: 8KQ reference .................................................................................. 147
Element C: 8KQ applicability ............................................................................. 149
Element D: analyzing individual KQ ................................................................. 151
Element E: weighing relevant factors and deciding .................................... 154
Intervention profile for class that demonstrated the most improvement over time .. 155
Summary............................................................................................................. 159
CHAPTER 5 ........................................................................................................ 162
Discussion ......................................................................................................... 162
Specific Implications for JMU ......................................................................... 162
Recommendations for modifications to the ER intervention checklist .......... 164
Recommendations for continued faculty development and support .......... 168
Strategies for scaling up the ER intervention at JMU ................................ 169
Study limitations............................................................................................... 173
Future Research Directions ........................................................................... 178
Beyond JMU: Broader Implications for Higher Education ......................... 18080
Appendix A ....................................................................................................... 212
Appendix B ....................................................................................................... 217
Appendix C ....................................................................................................... 218
Appendix E ....................................................................................................... 221
Appendix F ....................................................................................................... 222
Appendix G ....................................................................................................... 226
References ......................................................................................................... 227
List of Tables

Table 1. Summary of Data Management and Motivation Filtering Process Results ......184
Table 2. Demographic Information for Students Included in Fall 2016 Pre/Post Matched ER-WR Sample, Fall 2016 Pre/Post Matched ERIT Sample, Fall 2015 Comparison ER-WR and ERIT Samples, and Spring 2016 Comparison ER-WR and ERIT Samples ........185
Table 3. Average ER-WR Essay Scores by Element for Students Who Have Not Experienced Any Ethical Reasoning Interventions at JMU Compared to Students Who Have Experienced at Least One Ethical Reasoning Intervention at JMU .........................186
Table 4. Initial Known Groups Validity Evidence for ER-WR Scores ............................187
Table 5. Variance Components, Relative Reliability, and Absolute Reliability Estimates for Pre- and Post-test ER-WR scores for the (R*S)*E G-theory Design ..................188
Table 6. Variance Component for Object of Measurement, Relative Reliability, and Absolute Reliability Estimates for Pre-test ER-WR Scores Stratified by Rater Team for the (R*S*E)P G-theory Design .........................................................189
Table 7. Rater Harshness ER-WR Essay Ratings for Implant (“Common”) Essay #1 Used During ER-WR Rating Session on Tuesday, January 3rd .................................................190
Table 8. Rater Harshness ER-WR Essay Ratings for Implant (“Common”) Essay #2 Used During ER-WR Rating Session on Wednesday, January 4th .............................................191
Table 9. Raw and Statistical Comparison of Students’ ER-WR Scores at Pre-test and Post-test for Each Rubric Element and for Overall (Average) ER-WR Scores ...............192
Table 10. Cross-sectional Comparison of ER-WR Scores for Students Who Experienced It’s Complicated to Students Who Experienced the ER Intervention Piloted in the Current Study ....................................................................193
Table 11. Longitudinal Comparison of ER-WR Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty .................................................................194
Table 12. Statistical Significance and Practical Significance for Longitudinal Comparison of ER-WR Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty ......................................................195
Table 13. Longitudinal Comparison of ERIT Total Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty ..................................................196
Table 14. Statistical Tests of Significance and Practical Significance for Longitudinal Comparison of ERIT Total Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty ......................................................197
Table 15. Comparison of Current Study Results to Results from Good (2015) .................198
Table 16. Summary Information about Each of the Six Classes Included in the Current Study .................................................................................................................................199
Table 17. Implementation Fidelity Results for Duration, Student Responsiveness, Adherence, and Quality Aggregated Across All Faculty Members Stratified by MC ER Fidelity Checklist Overall Program Components ........................................200
Table 18. Implementation Fidelity Results for Each Faculty Member Stratified by MC ER Fidelity Checklist Specific Features ..................................................................................201
List of Figures

Figure 1. Graphical Organizer of Study Terms ..................................................202
Figure 2. Visualization and Description of the Steps in Fulcher and Colleagues (2014) Simple Model for Learning Improvement .................................................................203
Figure 3. Description and Visualization of the Process Used During the Summer Institute to Help Faculty Create an Ethical Reasoning Intervention and Fidelity Checklist ........204
Figure 4. Depiction of the Two Facet Nested Design for G-theory Analysis with Shaded Black Area to Represent Error, Dotted Lines to Represent the Object of Measurement, and Solid Lines to Represent the Two Facets or Sources of Systematic Error........205
Figure 5. Comparison of Students from Good (2015) Study to Student Comparison Group Used in Current Study to Student Treatment Group Used in Current Study.......206
Figure 6. Number of Total Intervention Specific Features Implemented by Each Faculty Member Plotted in Comparison with Their Students’ Average Post-test ER-WR Essay Scores ......................................................................................................................207
Figure 7. Students’ Average Post-Test ER-WR Total Scores Plotted Against Average Perceived Student Responsiveness and Quality Ratings for All Faculty Participants with Red Line Representing University Strategic Plan Goal for 2020 ................................................208
Figure 8. Faculty Participant Profiles of Students’ Average Post-Test ER-WR Total Scores, Average Perceived Student Responsiveness, and Quality Ratings with Red Line Representing University Strategic Plan Goal for 2020 ........................................209
Figure 9. Faculty Participant Profiles of Student ER-WR Average Scores, Average Perceived Student Responsiveness, and Average Quality for All Faculty Participants Categorized According to Class Type (Orange = General Education Class with Lower Level Students, Grey = Primarily Active Learning or Seminar Based, Non-Required Course with Service Learning Component and Upper Level Students, Purple = Primarily Lecture Based, Required Course for Major with Upper Level Students) with Red Line Representing University Strategic Plan Goal for 2020 ................................................210
Figure 10. Faculty Participant Profiles of Student Z-score ER-WR Average Scores, Z-score Average Perceived Student Responsiveness, Z-score Average Quality, and Z-score ...........................................................................................................................................211
Abstract

Examples of demonstrable student learning improvement in higher education are rare (Banta, Jones, & Black, 2009; Banta & Blaich, 2011). Perhaps because outcomes assessment practices are disconnected from pedagogy, curriculum, and learning improvement. Through partnership with the Madison Collaborative, the current study aimed to bridge this disconnect. Specifically, researchers applied implementation fidelity methodologies (O’Donnell, 2008) to an academic program, under the guiding framework of the Simple Model for Learning Improvement (Fulcher, Good, Coleman, & Smith, 2014). In doing so, researchers helped faculty create and elucidate an ethical reasoning educational intervention and accompanying fidelity checklist. Both were well-aligned with a University-level ethical reasoning performance assessment tool, the ER-WR. Implementation fidelity methodologies were applied within a diverse group of courses during the fall 2016 semester (e.g., courses for general education, major requirements, electives, etc.). Fidelity data indicated the extent to which the ethical reasoning intervention was implemented with high fidelity. Outcomes assessment data were collected and integrated with fidelity data to determine the effectiveness of the implemented ethical reasoning intervention. Results provided evidence of statistically and practically significant improvements in students’ ethical reasoning skills. In addition, results suggested that specific features of the ethical reasoning intervention positively influenced students’ ethical reasoning abilities. This study provides an example of how assessment practices can be effectively integrated with curriculum and pedagogy to demonstrate learning improvement.
CHAPTER 1

Introduction

In the late 1970s, Sony engineers released the first-ever portable music device, the Walkman (Adner, 2012). Several years later, MP3 music players were proposed as the next generation Walkman, extending Sony’s ownership rights to thousands of popular songs. However, without the widespread availability of high speed internet or systems for file storage and music management, the promising idea of MP3 music players dissipated (Adner, 2012).

Flash forward to 2001, more than twenty years later, Apple released its first ever MP3 portable music device with accompanying music storage and management system (i.e., the iPod and iTunes) (Adner, 2012). Although Apple was decades “late” to the portable music device market, it was the first to integrate all of the components (e.g., broadband internet availability, a storage device, and music management software) into a wildly popular whole. Ever since then, Apple has utterly dominated the portable music device market (Adner, 2012). Despite being well positioned, leaders from Sony and other technology companies missed a huge opportunity; they failed to realize how their music holdings and technology could fit into a coherent system.

Granted, this example is far afield of higher education. Nevertheless, the way Apple seamlessly integrated MP3 music files, broadband internet capabilities, and a companion music management system is analogous to the aims of the current research project. There have been several promising innovations in the main facets of higher education: assessment, teaching, and learning. However, higher education practitioners have struggled to integrate them. This study used an assessment methodology (i.e.,
implementation fidelity) to explore a learning intervention, under the guiding framework of a learning improvement model (i.e., the Simple Model). As suggested by Fulcher and colleagues (2014), joining assessment and learning components should help facilitate demonstrably improved learning for important, university-level student learning outcomes. For this dissertation, I focus on an ethical reasoning (ER) student learning outcome.

**Defining Key Terms**

Because *improvement, intervention, program theory*, and other related terms are used inconsistently in the higher education literature (Smith, Good, Fulcher, & Sanchez, 2015), I provide definitions at the outset. Several of these terms are associated with assessment practice specifically, while others are related to learning and education broadly. Figure 1 provides a depiction of the relationships among these terms and how the terms are related within the current study.

**Assessment related terms.** Higher education assessment practices are typically characterized by two main purposes: accountability and improvement (Ewell, 2009). That is, practitioners and higher education stakeholders often use assessment results for either accountability or improvement purposes. *Accountability purposes* are the processes driven by requirements for compliance and external requests from accreditors, legislatures, consumers, and others. *Improvement purposes* are more formative, internal efforts fueled by engaged and interested faculty and administration (Ewell, 2009). However, the difference between improvement and accountability is not always clean cut; many accreditors incorporate learning improvement into accountability requirements, for example.
I exclusively focus on assessment for improvement purposes. But *improvement* is a nebulous term used inconsistently and sometimes haphazardly in assessment literature (Smith, Good, Sanchez, & Fulcher, 2015). When I discuss *improvement*, I do so under the explicit definition provided by Fulcher, Good, Coleman, and Smith (2014, p. 4): “making a change to a program and then re-assessing to determine that the change positively influenced student learning.” To further explain this definition, “making changes to a program” includes re-designing course curricula, re-aligning course content or materials with pedagogical techniques and assessment tools, modifying course scaffolding, implementing new pedagogies or teaching techniques, and other efforts.

Within the context of assessment for improvement, I employ a critical – yet underused – assessment practice or methodology known as *implementation fidelity*. Implementation fidelity has been defined as “the degree to which a program model [educational intervention] is instituted as intended” (Dhillon, Darrow, & Meyers, 2015, p. 9). In other words, implementation fidelity data indicate to what extent the *delivered* educational intervention (e.g., pedagogies, curricula, etc.) differs from the *designed* or *planned* educational intervention (Gerstner & Finney, 2013; O’Donnell, 2008). Other names for implementation fidelity include enacted curriculum, program integrity, treatment integrity, compliance, clinical effectiveness, and adherence (Dhillon, Darrow, & Meyers, 2015; Mellard, 2010). Low fidelity of implementation (i.e., the delivered intervention differs drastically from the designed intervention) can drastically affect practitioners’ interpretations of assessment results. In Chapter Two, I explain the specific components of implementation fidelity and the importance of fidelity research.
Throughout this study, I use implementation fidelity research methodologies as a means to more systematically integrate assessment practices with learning improvement. That is, although implementation fidelity is categorized as an “assessment” concept in Figure 1, I believe the practice of implementation fidelity can help faculty better align their program theory with learning outcomes, pedagogies, and curricula. I also believe that implementation fidelity can serve as a “bridge” that more strongly connects assessment practice to student learning.

Implementation fidelity research has been conducted and applied in the medical field for years (Bond, 2000; Rogers, Eveland, & Klepper, 1977). Yet, it has been underused and undervalued by educational research, especially higher education (Berman & McLaughlin, 1976; Dhillon, Darrow, & Meyers, 2015). As a result of the 1980s assessment movement (Ewell, 2009), the majority of assessment practitioners have shifted their focus from teaching processes to learning outcomes. Thus, most practitioners typically do not feel the need to examine fidelity of implementation. They instead (mistakenly) infer the learning outcomes were achieved as a result of the planned educational interventions, not what actually happened in the classrooms when the interventions were implemented. Indeed, collection, analysis, and integration of implementation fidelity data is often missing from institution- and program-level assessment cycles.

In implementation fidelity research contexts, the *black box*, as referenced in Figure 1, is a commonly used term (McLaughlin, 1987). I use this term to represent the unknown intervention that occurs in classrooms and programs when fidelity data are not collected. More specifically, this term refers to the fact that it is unknown:
• if and what information, material, activities, or other specific features that comprise the educational intervention were actually delivered to students,
• whether specific intervention features were delivered for an appropriate length of time,
• whether specific intervention features were delivered with varying degrees of quality, or
• whether students were responsive enough during intervention delivery to actually internalize material.

Therefore, when assessment practitioners and faculty members analyze outcomes assessment data from multiple-choice tests, essays, or other measures, they do not know what educational intervention they are actually making inferences about. Is it the intervention as designed or is it some deviant of that? Bath, Smith, Stein, and Swann (2004) conclude that when practitioners attempt to “validate” or “quality assure” a curriculum, they must have consistency or homogeneity among the planned, enacted, and experienced curricula; an impossible feat absent implementation fidelity data.

As an example, imagine that several faculty members decided to create and adopt a new educational intervention to improve students’ ER abilities. However, the faculty do not collect implementation fidelity data; therefore, they have no information to suggest what actually occurred in the classroom when the educational intervention was supposed to be implemented. The delivered intervention is a black box. Inside this black box could be the intervention as it was designed or intended, or an intervention that severely deviated from what was intended. The black box obfuscates inferences from assessment data. Faculty cannot know whether the conclusions they are drawing from students’ scores are in reference to the intervention as they conceptualized and planned it, or in reference to some derivation of the planned intervention that was actually delivered to students.
Even if faculty collect implementation fidelity, they may still need didactic guidance concerning the use of assessment results to inform learning improvement. Therefore, I discuss and use a learning improvement model championed by Fulcher, Good, Coleman, and Smith (2014) that provides such guidance. The Simple Model (see Figure 2) describes a methodology for demonstrating improved student learning via integration of assessment, pedagogy, curriculum, and program theory. Although other improvement models existed prior to 2014 (Chaplot, Booth, & Johnstone, 2010), the Simple Model offers more didactic, practical guidance. In addition, Fulcher and colleagues (2014) provide an applied example of how the Simple Model can be used to guide demonstrable learning improvement. Other preexisting models tended to offer theories and strategies for promoting learning improvement, but less applied exposition or advisory guidance compared to the Simple Model.

According to Fulcher and colleagues’ (2014) Simple Model, improved student learning is demonstrated “when a re-assessment suggests greater learning proficiency than did the initial assessment” (p. 5). The model includes three core components:

1. Program faculty **assess** using sound instruments that tightly align with programmatic student learning objectives and directly measure student learning;

2. Assessment professionals and faculty development experts **intervene** at the faculty level by working with faculty to help them with course re-design, course scaffolding, pedagogy, and curriculum development skills. Through this faculty-level intervention, professionals help faculty modify existing educational interventions and/or create new educational interventions (e.g.,
pedagogies, curricula). Program faculty then \textit{intervene} at the student-level by implementing those new or modified educational interventions;

3. Program faculty \textbf{RE-assess} using the same instrumentation; and reassessment results indicate that post-intervention student learning actually improved compared to pre-intervention assessment results.

The three core components (e.g., assess, intervene, RE-assess) are further delineated into the two following steps:

1. \textbf{Readiness for Initiative}, which includes:
   a. Faculty nucleus dedicated to the learning improvement initiative
   b. Administrative support of the learning improvement initiative, and
   c. Rigorous assessment methodology involving a longitudinal data collection design.

2. \textbf{Planning Educational Intervention}, which includes:
   a. Identify one or two student learning outcomes (SLOs) to focus on,
   b. Investigate the current educational interventions already in place regarding the targeted SLOs and propose reasons students might not be achieving these SLOs,
   c. Propose learning modifications or create a new educational intervention, and
   d. Detail timetable for educational intervention to be implemented and assessments to take place.

When using the Simple Model to better integrate assessment practices with learning improvement, I refer back to these specific steps to situate the current research project
within the specific contexts of the Simple Model. For instance, implementation fidelity data can be used to address step 2b (*investigate the current educational interventions already in place regarding the targeted SLOs and propose reasons students might not be achieving these SLOs*). Unfortunately, prior to this research project, stakeholders had not gathered fidelity data related to ethical reasoning (ER) interventions. If they had, then these data could have helped address 2b and 2c.

For this research project, I collected fidelity data that can be used to address step 2b. The current research project also addressed steps 2c (*propose learning modifications or create a new educational intervention*) and 2d (*detail timetable for educational intervention to be implemented and assessments to take place*).

**Learning and education related terms.** In addition to assessment-related language, I use various terms associated with learning and education concepts. For instance, as depicted in Figure 1, I discuss *program theory* in the contexts of educational interventions and implementation fidelity. Bickman (1987) defined program theory as “a sensible model of *how a program is supposed to work*” (p. 5). He also cited Conrad and Miller (1987) in explaining program theory as “a system of beliefs, values, and goals that define the structure, process, and outcomes of a program” (p. 5).

More simplistically, an educational program (e.g., educational intervention) is designed to address an issue (e.g., ER skills) for a specific group of people (e.g., students). Given the program is well-designed and implemented with high fidelity, it should positively affect the targeted group (e.g., if a well-designed ER educational intervention is implemented with high fidelity it should help students enhance their ER abilities).
Program theory can also be defined as a means by which to show how an intermediate outcome can lead to a more comprehensive, ultimate outcome (Rogers, Petrosino, Huebner, & Hacsi, 2000); a variable can mediate the relationship between two other variables (Pedhazur, 1997, p. 159). For example, imagine that practitioners implemented an educational intervention intended to increase students’ ER skills from “Good” to “Excellent” as measured by a performance assessment rubric. Practitioners could assess whether or not the intervention increased students’ positive attitudes towards and perceived value of ER skills (as measured by an attitudinal assessment tool) and then assess whether or not valuing ER skills contributed to overall increases in students’ ER skills (as measured by an ER performance assessment rubric). In other words, is attitude toward ER skills a mediator variable between the educational intervention and students’ scores on the ER performance assessment rubric?

Program theory provides a model of how a given educational intervention is expected to work (Rogers, Petrosino, Huebner, & Hacsi, 2000). Expanding on Bickman’s (1987) conceptualizations, faculty members should create and articulate a program theory which details the specific aspects of their curriculum and how that curriculum is supposed to work – in theory – to enhance student learning, help students acquire a certain skillset, and more. The program theory is in reference to specific outcomes (i.e., criteria). Hence, program theory is related to learning outcomes assessment practice because it is the program theory that is being assessed or evaluated. That is, the purpose of conducting outcomes assessment is to understand if the educational intervention – which is operationalizing a clearly articulated theory of how students should acquire certain knowledge and skills – is effective.
I conceptualize program theory as the link between the *specific program features* (O’Donnell, 2008) of the intervention and the intended student learning outcomes. The program theory explains *why* and/or *how* certain specific features should result in certain student learning outcomes. To contextualize what program theory means, consider the following ethical reasoning example.

Student will do “X” activity, which evidence suggests should help them improve “Y” skill. For example, students will “review an example of “effective” or “good” ethical reasoning and describe the key characteristics of “good” ethical reasoning” to help them “apply ethical reasoning to their personal, professional, and civic lives.” Here, the activity that students are doing (i.e., “review an example of “effective” or “good” ethical reasoning and describe the key characteristics of “good” ethical reasoning”) is one example of a specific feature on the fidelity checklist. And the skill that students should improve (i.e., “apply ethical reasoning to their personal, professional, and civic lives”) is one of the Madison Collaborative’s Student Learning Outcomes. According to the program theory, when students *review an example of “good” ethical reasoning and describe the key characteristics of “good” ethical reasoning*, their abilities to *apply ethical reasoning to their personal, professional, and civic lives* should improve because the activity:

- provides students with a concrete example of what “good” ethical reasoning looks like. Cognitive research suggests that providing students with high quality examples can promote comprehension and retention;

- forces students to become familiar with the components or characteristics of “good” ethical reasoning skills. Educational research suggests that increasing
students’ familiarity with concepts—through quizzing, assignments, demonstrations, exams, etc. – can promote retention;

- makes students describe “good” ethical reasoning skills via their own conceptualizations or understandings. Cognitive research suggests that having students describe concepts using their own understandings, language, metaphors, etc. can help them internalize and solidify information. Then students can more readily apply the information to their lives.

As shown in the ethical reasoning example, the program theory explains why having students do the specific program feature (i.e., review an example of “effective” or “good” ethical reasoning and describe the key characteristics of “good” ethical reasoning) should result in students achieving the specific learning outcome (i.e., apply ethical reasoning to their personal, professional, and civic lives). The “should” justification emanates from cognitive and educational theories, which suggest that certain type of activities should contribute to intended learning outcomes.

By linking program theory with implementation fidelity, I aimed to better integrate assessment practice with learning improvement (see Figure 1). Implementation fidelity helps faculty members better articulate their program theory and subsequently align that theory with learning outcomes, specific features of educational interventions, and assessments. If faculty can understand that implementation fidelity is connected to program theory, perhaps they will also be able to see that assessment processes (which can include implementation fidelity) can be integrated with tenets of learning and education (which include program theory).
As Figure 1 shows, program theory should inform the activities, programming, exercises, etc. that make up educational interventions. These educational interventions can occur at two different, yet connected, levels: student and faculty. I discuss both levels of educational interventions in relation to improving students’ abilities to apply ethical reasoning, which is the substantive area of interest for the current research project.

First, a student-level educational intervention involves pedagogy and/or curricula implemented to positively affect students’ learning. Student-level educational interventions can encompass activities, demonstrations, examples, assignments, discussions, case studies, projects, papers, feedback, etc. that occur both in and outside of the classroom. More specifically, student-level interventions represent all of the experiences that faculty provide to students with the aim of enhancing students’ knowledge, thinking, and skills. Such experiences should help students achieve intended learning outcomes given cognitive, motivational, behavioral, and other theories and research. Indeed, faculty would have selected or created the experiences that they provide to students based on various theories of learning, retention, retrieval, and so forth.

For example, consider faculty members delivering a class using a different pedagogical style, like interteaching (Boyce & Hineline, 2002; Saville, Zinn, Neff, Normam, & Ferreri, 2006), to help improve students’ ethical reasoning skills. Use of interteaching would constitute a student-level educational intervention. In this example, interteaching (i.e., the intervention) should help students achieve the intended learning outcomes (i.e., the SLO) because previous research has shown that this type of pedagogical strategy enhances students’ abilities to think analytically and apply information in real-world settings (i.e., the program theory). Said another way, the use of
the interteaching pedagogy is theory-based (i.e., there is a program theory). Previous research indicates why interteaching should improve students’ abilities to apply ethical reasoning skills.

A faculty-level educational intervention involves development, training, and/or continuing education opportunities. The purpose of development and training is to help faculty effectively implement educational interventions. Faculty-level interventions could emphasize delivering a re-designed curriculum or using a more appropriate teaching strategy, for example.

Perhaps faculty members completed a week-long workshop that taught them how to re-design and deliver their classes using interteaching pedagogies. The workshop would constitute a faculty-level intervention. Faculty learned about the theory underlying interteaching and how interteaching pedagogies or activities can align with their program theory. As a result of the workshop, faculty had the tools to implement interteaching pedagogies and understand why interteaching should help students achieve learning outcomes.

Both levels of educational interventions can be further broken down into pedagogy and curriculum (See Figure 1). For the current research, pedagogy refers to various teaching strategies or techniques that can be used in the classroom. Cognitive scientists have found that certain, evidence-based pedagogical techniques can promote learning and retention (Halpern & Hakel, 2003).

Interteaching is one example of an evidence-based pedagogical technique that includes elements of cooperative learning, problem-based learning, and reciprocal peer-tutoring (Boyce & Hineline, 2002; Saville, Zinn, Neff, Normam, & Ferreri, 2006).
Interteaching requires faculty to create a “prep guide” that directs students through specific material before they come to class. Students answer the prep guide questions individually before class and then work through the questions as a group during class (e.g., reciprocal peer-tutoring). Rather than lecturing, the faculty member spends class time interacting with students, answering questions, and facilitating group discussions (Saville, Zinn, Neff, Normam, & Ferreri, 2006). At the end of class, students indicate which questions or concepts were the most difficult. Then the faculty member creates a short lecture to address difficult concepts during the next class session (Saville, Zinn, Neff, Normam, & Ferreri, 2006). Other examples of evidence-based pedagogies include:

- inquiry-based teaching (based on constructivist learning theories put forth by psychologists Piaget, Dewey, Vygotsky, and others), and
- flipped classrooms (based on blended learning and learner-centered education theories), active learning (based on constructivism, cooperative learning, and learner-centered education theories).

Without adequate development and training opportunities, faculty may be unaware of evidence-based pedagogical techniques, when to use them, and how to appropriately implement them into classes. To integrate assessment practices and learning improvement, institutions must provide adequate faculty development opportunities related to pedagogy.

Recall, educational interventions are comprised of both pedagogy and curriculum (See Figure 1). *Curriculum* is a term that can take on various meanings. The curriculum typically encompasses various activities, assignments, demonstrations, projects, etc. that take place in and outside of the classroom. The curriculum should provide opportunities
for students to learn concepts and practice skills that they need to achieve intended student learning outcomes. For my purposes, curriculum is synonymous with intervention specific features (O’Donnell, 2008).

Cognitive research suggests that when students are asked to complete certain types of activities, learning can be enhanced (Halpern & Hakel, 2003). Consider faculty who implement a classroom activity during which students have to draw concept maps to visually represent information they heard during a presentation. Such an activity can help students process and encode information in multiple formats, including visuospatial and auditory. Multiple format encoding activities can promote information recall better, compared to single format encoding of information (Halpern & Hakel, 2003, p. 39).

Now, consider a more concrete example, within the contexts of this study. Imagine that faculty members included an activity (i.e., an intervention specific feature), within their curricula, that required students to:

- take a code of ethical standards that was originally presented textually,
- re-present the standards graphically or visually, and
- visually incorporate the 8KQ into the graphic.

According to Halpern and Hakel (2003), such an activity could enhance students’ abilities to recall the 8KQ. The activity would be part of the curriculum (i.e., an intervention specific feature on the checklist). And the activity would contribute to program theory. That is, the underlying program theory could be conceptualized as follows: Cognitive research suggests that when students have to re-present information in new forms, learning and retention are enhanced. Faculty implemented an activity that requires students to take ethical standards presented textually and re-present them
visually, using the 8KQ. Given cognitive theories, the “re-present” activity *should* improve students’ abilities to recall the 8KQ and perhaps subsequently apply their ethical reasoning skills.

Faculty can successfully re-design curricula, align curricula with program-level assessment instruments, and scaffold curricular components across a course or program. But often, faculty need development opportunities (i.e., faculty-level educational interventions) to learn how to do so. Therefore, it is important to provide adequate development opportunities to help faculty design effective, evidence-based curricula.

In addition to pedagogy and curriculum, an important component of faculty-level educational interventions is faculty development (see Figure 1). Faculty development is an important term to define because a broad range of activities are often considered “development opportunities.” For instance, some consider an assessment center providing guidance on data collection methodology a form of faculty development. However, for this study, *faculty development* exclusively refers to faculty training or education directly related to educational interventions (i.e., pedagogy and curriculum). In other words, the term faculty development is used to refer to opportunities that directly influence or enhance faculty members’ teaching abilities. Faculty development can include opportunities to learn about and apply course re-design, course alignment, course scaffolding, pedagogical techniques, and learning theories. The following are more concrete examples of faculty development experiences:

- faculty participate in a week-long institute on course re-design,
- faculty complete training from teaching experts about how to apply various pedagogical strategies within certain disciplinary contexts, or
• faculty attend a workshop where they learn how to create a fidelity checklist and use it to better align educational interventions with learning outcomes and assessment instruments.

Statement of Problem

Since the 1980s, the field of assessment has progressed substantially, promoting accountability and rigorous methodologies (Fulcher, Smith, Sanchez, & Ames, 2017). Increased attention on implementation fidelity and meta-assessment (Fulcher & Orem, 2010) contributed to innovations in assessment practices. Similarly, conversations touting the benefits of High Impact Practices (HIPs) (e.g., learning communities, undergraduate research, writing-intensive courses, etc.) have had a positive effect, moving faculty toward conversations of teaching and learning (Kuh, 2008). It has been beneficial for higher education stakeholders to discuss the kinds of educational interventions that should, theoretically, have a high impact on students’ learning.

Conversations about educational interventions are fortuitously occurring concomitantly with learning improvement efforts. For instance, there have been various national-level initiatives concerning use of assessment results for learning improvement (Blaich & Wise, 2011) and cultures of inquiry around student learning improvement (Chaplot, Booth, & Johnstone, 2010). With the introduction of the Simple Model (Fulcher, Good, Coleman, & Smith, 2014), practitioners gained a guiding framework to move beyond mere conversations about if and why particular educational interventions may work, to empirically evaluating interventions in real-world learning environments. That is, the Simple Model provides a framework or mechanism by which researchers can
empirically examine whether various educational interventions “work” (i.e., improve student learning) as hypothesized.

No doubt these advances have benefited higher education, but their effects have been less than one might expect, especially as they relate to improvement. A disconnect persists among assessment practice, pedagogy, curriculum, and learning improvement. And that disconnect incites questions regarding the worth of higher education (Arum & Roksa, 2011; Taylor et al., 2011).

Much like Sony and other technology companies in the 1970s, higher education practitioners have not fully integrated all of the appropriate components for demonstrating learning improvement, until now. Practitioners have failed to coherently put all of the components together, including:

- articulating sound program theory,
- aligning interventions with learning outcomes and theory-based pedagogies,
- appropriately modifying pedagogies and curricula,
- re-assessing after modifications are implemented, and
- examining intervention implementation fidelity.

With the introduction of the Simple Model and applications of implementation fidelity research to educational contexts, the necessary components are in place. Now, through the current research, I proceeded as Apple did in the early 2000s. I assembled the necessary components (e.g., the Simple Model for learning improvement, implementation fidelity research, etc.) as a coherent, integrated whole to demonstrably improve student learning.
CHAPTER 2

Literature Review

Purpose of Literature Review

The purpose of this literature review is two-fold: to explore the disconnect between student learning outcomes assessment and student learning improvement, and to offer a solution by integrating the Simple Model (Fulcher, Good, Coleman, & Smith, 2014) and implementation fidelity research (O’Donnell, 2008). First, it is important to consider the historical contexts and the current state of affairs concerning higher education assessment practices, focusing on significant methodological advancements. Nevertheless, despite these achievements, demonstrations of learning improvement remain elusive. Thus, I describe several underlying causes and a solution for this issue.

An essential feature of student learning improvement is effective learning interventions, implemented with high fidelity. I describe the components of implementation fidelity, examine fidelity research in educational settings, and convey the need for more fidelity research in higher education academic contexts. To better integrate assessment practices with learning improvement, the Simple Model and implementation fidelity research are applied to a specific substantive content area (i.e., a university-level ethical reasoning education program) to address four main research questions.

Historical Contexts of Assessment Practice

Research examining student learning and development in higher education was first documented in the 1930s (Ewell, 2002). Approximately thirty years later, in the 1960s, a growing body of research related to student learning and development existed; enough to prompt the assessment movement ten years later (Ewell, 2002; Feldman &
Newcomb, 1969). Additionally, the 1960s and 70s saw program evaluation arise in higher education contexts. The juncture of these two developments formed the basis of student learning outcomes assessment.

By the 1980s, stakeholders from both inside and outside of higher education were calling for reform and accountability. From within, organizations stressed the need for improved general education curricula that were clearly articulated and able to be assessed or evaluated. Moreover, faculty “ought to be willing to engage in assessment” as part of their scholarly work (Ewell, 2002, p. 8). External stakeholders, like state governments, wanted to hold institutions increasingly accountable for their use of taxpayer dollars and return on investment for graduates. State-level legislators and stakeholders saw student learning outcomes assessment as one means to achieve such accountability.

However, state-level mandates related to assessing student learning and development experienced limited success; state budgets did not allow sufficient funds to enforce assessment-related mandates (Ewell, 2009). Furthermore, state-level mandates emphasized comparability and thus did not allow institutions enough flexibility to establish their own learning outcomes and methods of evidencing learning growth or development (Ewell, 2009). With the establishment of regional accrediting bodies in the 1990s, states could drop their (mostly) ineffective mandates and transition accountability responsibilities to accreditors. Therefore, accreditors became a “buffer” between institutions and federal- and state-level governing bodies (Ewell, 2009).

Unlike state-level legislations, accreditors allowed institutions the necessary flexibility to articulate their own learning outcomes, expectations for students, assessment tools, assessment methodologies, etc. Today, regional accreditors remain a key
stakeholder in higher education assessment, serving as “gatekeepers” for access to federal grants, loans, and other important student financial aid sources. Their charge: to keep institutions accountable for student learning and development, while also ensuring quality and continuous improvements. A key hypothesis underlying the role of accreditors was that ensuring institutions participate in high quality assessment practices should eventuate in “better” or improved student learning. As discussed below, however, this hypothesis has been called into question.

Indeed, since the 1980s, assessment practices have improved noticeably. Practitioners have successfully refined their methodologies to meet accountability demands and demonstrate value (Ewell, 2009), which may (in part) be related to the role regional accreditors have played. For example, over the past ten years, academic programs at James Madison University (JMU) have demonstrably improved the quality of their assessment practices (Rodgers, Grays, Fulcher, & Jurich, 2013). Institutions across the United States are also engaging in more assessment practice and of higher quality (Kuh, Jankowski, Ikenberry, & Kinzie, 2014). According to Kuh and Ikenberry (2009) and Kuh et al. (2014), approximately 84% of institutions nationwide have articulated a common set of student learning outcomes that apply to all students, have implemented bottom-up assessment initiatives fueled by commitment to learning improvement, and have reported substantial perceived support for assessment practices on campus.

In other words, institutional stakeholders are investing more than just time and interest into assessment processes. As of 2013, on average, institutions spent approximately $160,000 annually on assessment activities (Terrell, 2013). And over 71%
of surveyed assessment practitioners agreed that the benefits of engaging in assessment processes outweighed these costs (Terrell, 2013).

This increasing investment and advancement in assessment practice should contribute to better student learning, in theory. Yet, higher education examples of learning improvement are exceptionally rare. The following section explicates the disconnect between assessment practices and learning improvement.

**Better assessment ≠ Improved learning.** Using assessment results to influence educational interventions and subsequently demonstrate learning improvement has been part of national discussions for decades. Encouragingly, “commitment to improvement” was reported as one of the “top three most influential forces driving assessment practice” in American institutions (Kuh & Ikenberry, 2009, p. 7). These results suggest that assessment practitioners’ intentions were appropriately aligned with improvement. Many believed that improved assessment quality (e.g., more rigorous methodology, higher quality data, more easily interpretable reports, etc.) would “inevitability lead to improvements in student learning” (Blaich & Wise, 2011, p. 8). Conflation of assessment practice with improvement may explain why higher education has disproportionately focused efforts on conducting research and disseminating practices concerning the methodologies of assessment, rather than focusing on improvement per se.

Nonetheless, recent studies suggest the link between high quality assessment and improved learning is not strong. For instance, Banta, Jones, and Black (2009) reviewed 146 exemplary assessment reports from institutions across the United States in search of improved student learning. Such examples were found in only 6% of these exemplary assessment reports (Banta, Jones, & Black, 2009; Banta & Blaich, 2011). A few years
later, Blaich and Wise (2011) lead a national study supported by the Center of Inquiry at Wabash College. They tested the hypothesis that “a lack of high-quality data was the primary obstacle that institutions faced in using assessment evidence to promote improvements in student learning” (2011, p. 8). They dismissed the hypothesis. Blaich and Wise recognized that the endeavor overemphasized methodology and data analysis. Little time and resources were spent helping institutions use assessment data in a meaningful and intentional way to improve learning. Furthermore, they realized that while many institutions collected high quality, actionable assessment data, few actually used the data to influence change and evidence improved student learning (Blaich & Wise, 2011).

Nationally, leaders have voiced other concerns with current assessment practices; as Suskie (2010) observed, “…today we seem to be devoting more time, money, thought, and effort to assessment than to helping faculty help students learn as effectively as possible” (“Why Are We Assessing?,” para. 8). It is true that higher education institutions across the United States appear to be increasing the frequency with which they use assessment results for accreditation, strategic planning, program review, and curricular modification; yet, higher education assessment practitioners are realizing that these efforts may not be enough to actually improve student learning in a demonstrable way (Kuh & Ikenberry, 2009; Kuh et al., 2014).

Practitioners and stakeholders from my institution’s assessment office have come to a similar realization. At James Madison University (JMU), academic and student affairs programs have struggled to parlay excellent assessment practice into demonstrably improved student learning, despite well-resourced assessment support. For instance, JMU
has a nationally awarded meta-assessment system (CHEA, 2015) and the nation’s largest student learning outcomes assessment center (the Center for Assessment and Research Studies). Similar to Blaich and Wise’s (2011) findings, colleagues at JMU discovered that high quality assessment practices and methodologies rarely yielded student learning improvement.

The original hypothesis for this disconnect was that faculty were not responding to results. Indeed, if programs do nothing differently after examining results (i.e., applying new educational interventions) one would not expect learning to improve. However, findings from a recent study investigating use of results (Fulcher, Smith, Sanchez, & Sanders, in press, 2017) suggested an alternative hypothesis.

Assessment specialists from JMU’s Center for Assessment and Research Studies (CARS) coded assessment reports from specific academic degree granting programs (Fulcher, Smith, Sanchez, & Sanders, in press, 2017). The academic degree programs included in the study had received “exemplary” ratings for the quality of their assessment practices based on a meta-assessment rubric (Fulcher & Orem, 2010). Encouragingly, Fulcher, Smith, Sanchez, and Sanders (in press, 2017) found that faculty reported implementing learning interventions that were typically coded as “moderately strong” and driven by data from direct assessment measures (i.e., data from previous assessments that used direct measures). In other words – as opposed to our initial hypothesis – faculty were taking action on results. Yet, the exact components or specific features of the educational interventions were not elaborated. In addition, programs rarely reported conducting follow-up re-assessment to determine whether changes to pedagogy,
curriculum, or educational interventions actually improved student learning outcomes (Fulcher, Smith, Sanchez, & Sanders, in press, 2017).

Furthermore, programs did not report data regarding the fidelity with which the educational interventions were implemented. That is, assessment reports included no data indicating the extent to which the delivered educational intervention deviated from the planned or intended intervention (Gerstner & Finney, 2013; O’Donnell, 2008). Granted, faculty were not asked to collect or report implementation fidelity data. But without such data, faculty cannot know if the intervention they are making inferences about, based on students’ outcomes assessment data, is the intervention as described on paper or some derivation thereof. Thus, making intentional and informed changes to the intervention that will eventuate in demonstrably improved learning is nearly impossible.

Perhaps this lack of intervention specificity, re-assessment, and implementation fidelity data have contributed to an overall inability to demonstrate student-learning improvement. Clearly, rigorous or “quality” practices (e.g., sound instrumentation, longitudinal designs, etc.) have been beneficial to progressing the field of educational assessment. But practitioners have not been able to effectively bridge the gap between assessment practices and student learning improvement. The disconnect echoes findings from the aforementioned national research studies (Banta, Jones, & Black, 2009; Blaich & Wise, 2011). Practitioners will not be able to effectively integrate assessment practices with learning improvement until they have a better understanding of the existing disconnect.
Exploring the Disconnect Between Assessment Practices and Learning Improvement

As discussed previously, one of the main purposes for conducting assessment is for learning improvement (Ewell, 2009). Why then is there a disconnect between the practice of assessment and demonstrably improving student learning? Why aren’t curriculum, pedagogy, and assessment collaboratively guiding instruction and learning?

Given observations from assessment practices at my own institution, I explored several reasons for this problem. Culprits include: the level problem, unclear and inconsistent communication, lack of easily accessible and detailed learning improvement examples, general culture issues, and lack of focus on program theory and educational interventions.

The Level Problem. Assessment initiatives at most higher education institutions tend to occur at the program level. Meanwhile, faculty-level interventions or development opportunities typically occur at the individual faculty member level. As described by Good (2015), this creates a level problem that makes it nearly impossible to align program-level assessment with faculty training or development centered around individual instructor’s courses:

While faculty development initiatives tend to focus on individual sections of courses, program assessment is focused at the academic program level. There is a notable disjunction between the two. Redesigns of individual courses are valuable for each professor who engages in the process and are likely beneficial for his or her students as well. However, when a program-level weakness in student learning is discovered, rarely is the solution found in a single section. Typically,
multiple sections of the same course and/or sequences of courses are in question… For example, if a program is concerned about graduates’ ability to analyze data, faculty members may need to coordinate an intervention involving a sequence of courses. Departments rarely have the time, expertise, or motivation to coordinate such a complex effort. Thus, an intervention that infuses sound faculty development principles (e.g., course design and learner-centered approaches) is needed at the program level to create systematic strategies that will improve student learning. (p. 30-31)

Faculty at JMU – similar to most higher education public institutions – typically implement educational interventions within their individual class sections. But assessment for their department occurs at the program-level. Thus, it is difficult for faculty to determine how they can use results from program-level assessments to inform pedagogical and curricular changes within their individual classes or sections.

Imagine a hypothetical faculty member teaching in a nursing program who wants to improve students’ abilities to apply ethical reasoning to their professional lives. She recognizes that she lacks expertise in pedagogy, curriculum, and program theory; therefore, she cannot appropriately effectively modify her course relative to the specified learning outcome. Fortunately, her university has a faculty development center. There, she spends the summer consulting with faculty developers who help her articulate her program theory and align that with an appropriate theory-based pedagogical technique (e.g., interteaching), course assignments, in-class activities, and her ethical reasoning student learning outcome. Faculty developers also help her re-design the course such that the specific features of the educational intervention she delivers to students should result
in the intended learning outcome given the cognitive learning theories that comprise her program theory. At the beginning of the fall semester, she implements the changes in her class (e.g., uses interteaching, uses modified class activities, etc.) with high fidelity, in the hopes of improving students’ ethical reasoning skills.

All students in the nursing department are required to complete assessments at the beginning and end of the semester. One of the assessment tests is an ethical reasoning assessment that is aligned with the ethical reasoning student learning outcome that the faculty member incorporated into her course. During a faculty meeting, at the end of the semester, her department head shares the aggregated (program-level) assessment results for the ethical reasoning test. On average, students ethical reasoning skills either stagnated or declined over the course of the semester (i.e., from pre-test at beginning of semester to post-test at end of semester). The faculty member is confused by these results; she felt that her re-designed course should have helped her students better achieve the ethical reasoning student learning outcome. She wonders how she could use these program-level assessment results to further modify her teaching strategy or class activities. Given the results are at the aggregated program-level, she finds no guidance for modifications for her individual course. Frustrated and discouraged, she abandons the ethical reasoning outcome and discontinues teaching ethical reasoning in her course.

In this hypothetical example, curriculum and pedagogy experts collaborated with an individual faculty member to articulate a program theory, redesign a course section accordingly, and apply evidence-based pedagogies. However, subsequent program-level assessment results offered little guidance for her course. Indeed, there was a level
problem between her teaching and learning efforts in one individual course section and assessment across many courses within the program.

To alleviate this level problem, the classroom level modifications needed to be made across various classes within the program, not simply in one isolated section. That is, the entire program faculty needed to collaboratively work with faculty developers to make intentional and theory-driven modifications across a range of courses within the program. It is expected that the efficacy of modifications made to just one isolated course or section level (even if those modifications were theory-driven and effective in improving learning), will not come through or be evidenced in assessments that occur at the program level. Ultimately, both faculty development and assessment need to be applied at the program-level to overcome the level problem.

**Unclear and inconsistent communication.** Another potential explanation for the disconnect is unclear and inconsistent communication among faculty and assessment professionals, among accreditors and assessment professionals, and even among expert assessment professionals. For instance, faculty and assessment practitioners typically seek information about quality assessment practices from books, rubrics, and standards for best practices. But when these resources do not clearly and consistently communicate how assessment results can be *used* to influence and evidence learning improvement, is it a surprise that practitioners struggle to bridge the assessment-improvement gap?

Based on a qualitative review of popular assessment books, meta-assessment rubrics, and regional accreditation standards, research suggested that most definitions of “use of results to improve student learning” were vague and lacked detailed examples (Smith, Good, Fulcher, & Sanchez, 2015). For instance, only three of the fourteen books
reviewed mentioned the role of re-assessment in determining the success of educational interventions (Smith, Good, Fulcher, & Sanchez, 2015).

Prior to the dissemination of Fulcher et al.’s (2014) Simple Model, assessment practitioners did not have an integrated model or framework that didactically explained a process for evidencing student learning improvement. In the past, when various assessment experts attempted to explain such processes in their publications, rubrics, standards, etc. the messages were mixed, at best. A history of ambiguous and inconsistent communication may help explain why there are so few examples of demonstrable learning improvement in higher education today (Smith, Good, Sanchez, & Fulcher, 2015).

**Lack of easily accessible and detailed learning improvement examples.**

Research suggests that learning by example can facilitate conceptual understanding (Atkinson, Derry, Renkl, & Wortham, 2000; Bourne, Goldstein, & Link, 1964). Therefore, the lack of concrete, detailed examples that integrate assessment practices with demonstrable student learning improvement is cause for concern (Smith, Good, Fulcher, & Sanchez, 2015). Although some assessment resources like Suskie’s (2010) *Assessing Student Learning: A Common Sense Guide* include general examples of what it means to use assessment results for improvement, these examples may not be easily accessible given they are often buried within sections or book chapters.

Fulcher, Good, Coleman, and Smith (2014) published an article that was solely dedicated to explaining a process for demonstrable learning improvement (i.e., The Simple Model). Fulcher and colleagues included a detailed, hypothetical example of an academic program that used assessment results to evidence learning improvement. It was
important to provide readers with tangible examples to concretize and demystify how assessment practice can be integrated with learning improvement (Fulcher, Good, Coleman, & Smith, 2014). As more assessment practitioners reference and use detailed examples like those provided by Fulcher and colleagues, perhaps assessment and learning improvement will become clearer, more refined, and better connected.

**General culture issues.** Since the mid-1990s, scholars have urged higher education to transition from an “instruction-centered” (i.e., “the mission of higher education is to deliver instruction,” for example, via passive lecture-discussion formats where faculty talk and students listen) to a more “learning-centered” (i.e., “produce learning with every student by whatever means works best;” lecture is replaced with whatever approach best facilitates learning of specific knowledge by specific students) paradigm or system (Barr & Tagg, p. 13-14, 1995; Lumina Foundation, 2016). But the general infrastructure and reward systems of most higher education institutions make systemic change and innovation both sluggish and challenging (Spence, 2001). Not to mention, within higher education academic departments, “change” is not necessarily a feasible or desirable option (Knight, 1995). These cultural issues can create problems when programs attempt to use assessment practices to influence educational interventions (e.g., changes to pedagogies, curricula, etc.) and implement re-assessment processes to determine whether student learning actually improved.

For example, consider faculty members who start a learning improvement initiative to help students improve a specific skill important to their field. Faculty wanted to improve students’ abilities to apply the 8 Key Questions of ethical reasoning to their
personal, professional, and civic lives. Imagine that, to demonstrably improve students’ abilities to apply ethical reasoning, the faculty had to:

- reflect on their assessment results,
- elaborate what high quality ethical reasoning application behaviors looks like,
- examine the “as is” curriculum to understand where students are (and are not) experiencing opportunities to learn the skills to apply ethical reasoning,
- create a “to be” curriculum that appropriately scaffolds opportunities to learn ethical reasoning application skills across courses/experiences throughout the program
- learn about evidenced-based pedagogical techniques that are well-suited to teach ethical reasoning application skills and incorporate these into their classroom pedagogy
- construct a psychometrically sound assessment tool to measure their students’ ethical reasoning application skills, and
- collect and analyze pre and post test data using the assessment tool.

Now, imagine that the faculty received no additional funding to complete this work. Furthermore, this work was not included in their evaluations or counted towards their promotion and/or tenure pursuits. Even if faculty were intrinsically interested in such a process, it is doubtful they would have the time or determination to even attempt such a daunting learning improvement project. Other demands such as scholarship – which is more explicitly reinforced by their department – would likely receive attention instead.

Integrating assessment practices with student learning improvement requires an immense amount of time and coordination from program faculty. If faculty do not think they can demonstrate learning improvement or they do not value it, then they will not attempt such an endeavor. Moreover, when learning improvement is not appropriately incorporated into faculty members’ position responsibilities and is not rewarded via review or tenure structures, the culture around learning improvement will be lackluster at best. Cultural issues are a hindrance to more “learning-centered” higher education
systems (Barr & Tagg, 1995), and they contribute to the disconnect between assessment and learning improvement.

**Lack of focus on program theory and educational interventions.** Over the years, practitioners have focused less attention on program theory compared to other aspects of program evaluation, like assessment (Bickman, 1987). Inattention to program theory may be due to cultural issues, lack of adequate time, undervaluing of faculty work related to learning improvement initiatives, and so forth. In addition, the foundations of the assessment movement tended to be more focused on accountability than improvement (Ewell, 2002).

If faculty are not educated in learning theories (e.g., cognition, motivation, behavior, etc.) as part of their formal training, then they must take time to learn them through other means. Knowledge of learning theories is necessary to create evidence-based curriculum and implement evidence-based pedagogies. However, faculty are often unfamiliar with theories that inform evidence-based teaching practices. Furthermore, given the cultural issues discussed previously, faculty are rarely given the opportunity to learn about and incorporate theory-driven practices. Doing so would take time away from other duties and obligations. Why then would we expect students to receive effective learning interventions when faculty do not have time or support to complete appropriate training opportunities related to teaching and learning?

To better align assessment and improvement efforts and demonstrably improve student learning, practitioners must progress their assessment practice beyond just methodological considerations. Stakeholders must prioritize the role of theoretically-based educational interventions, which relates to the previous discussion concerning
cultural issues. That is, faculty work related to learning improvement needs to be valued and rewarded. Moreover, educational interventions should reflect or support program theory. But too often educational interventions are developed (and assessed) without a clear, empirically-studied theory as a foundation (Bickman, 1987). In the language of the Simple Model, too little time has been spent “feeding the pig” (i.e., intentionally modifying educational interventions, informed by program theory, and implementing them with high fidelity).

For many, the importance of student-level interventions is obvious. Indeed, students must engage with better learning environments (e.g., those that include evidence-based pedagogies, well-aligned curriculum, well-scaffolded courses, etc.) to enhance their learning and subsequently demonstrate learning improvement. But faculty-level interventions have not received as much attention. Learning support at the faculty-level, however, is no less important. How will students receive more potent learning environments without faculty development? As Spence (2001) describes, “we won’t meet the need for more and better higher education until professors become designers of learning experiences and not teachers” (p. 18).

Becoming better designers of student learning experiences does not happen by accident, however. Similar to students, faculty must receive interventions to become more competent designers of curriculum and deliverers of pedagogy. They must receive guidance regarding how to do the following effectively:

- assemble their program theory according to literature and best practices.
- clearly articulate the components or features of their program theory to students and other stakeholders.
These components or features, once articulated, can help faculty ensure that the student-level interventions they implement in their classrooms are well-aligned with pedagogical techniques, assessment tools, and specific learning outcomes they intend for students to achieve.

In the past, assessment practitioners have not explicitly included forms of faculty development and/or training in their assessment cycles. The cycles naively assume that once faculty reach the “Use of Results” stage, they will coordinate program-level efforts, create an effective intervention, make evidence-based changes, (wave a magic wand) and learning improvement will automatically eventuate. In reality, most program faculty are ill-equipped to proceed successfully. They may require further training or consultation from faculty developers who possess expertise in pedagogy and curriculum.

Imagine a faculty member who aspires for students to apply ethical reasoning skills. However, she does not have experience with effective ethical reasoning interventions. The faculty member would not know how to incorporate strategies (e.g., interteaching, meta-cognitive checks, inquiry based teaching, etc.) that were better aligned with the student learning outcome (e.g., students will apply the 8 Key Questions of ethical reasoning to their personal, professional, and civic lives). Or imagine the faculty member did not possess knowledge and experiences related to course re-design or course scaffolding. Clearly, this faculty member would struggle to change their pedagogy and curricula in a meaningful way. Student learning would not demonstrably improve.

Fortunately, faculty-level interventions are receiving more attention, especially as institutions pivot their focus toward student learning (Brancato, 2003; Gillepsie &
Robertson, 2010). But more focus on faculty-level educational interventions and program theory is still needed.

In addition, recall that assessment practice over the past thirty-five years has emphasized methodological rigor. Such a narrow methodological focus was likely necessary when assessment practice was in its infancy. However, such myopia may have deterred progress toward an integrated model of assessment practice and learning improvement.

Fortunately, the field of assessment has started to shift attention toward learning improvement. For instance, national organizations like the Association of American Colleges and Universities (AAC&U) have strived to pivot higher education toward a more “learning-centered” system via resources, conferences, publications (Barr & Tagg, 1995). Leadership and staff at AAC&U clearly understand the need for greater focus on program theory and educational interventions; this is demonstrated through their numerous case studies, conferences, workshops, online resources, and research related to education, integrative learning, faculty development and training, and so forth.

As mentioned previously, AAC&U’s dissemination and promotion of High Impact Practices (HIPs) have helped practitioners and faculty rethink the kinds of learning experiences that should comprise educational interventions (Kuh, 2008). Additionally, AAC&U has emphasized the important role of assessment. In 2009, they released the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics to assess student work across various disciplines and institutions. More recently, AAC&U – with the State Higher Education Executive Officers Association (SHEEO) – created the Multi-State Collaborative (MSC) to gather student artifacts from participating
institutions, evaluate them using the VALUE rubrics, and share results back to individual institutions. These efforts emphasize the importance of helping institutions understand how to best use the VALUE assessment rubrics, report results, and use information in a meaningful way.

Despite laudable efforts, AAC&U stakeholders have not tightly aligned their HIPs with explicit Student Learning Outcomes (SLOs). As a result, the potency or effectiveness of educational interventions and program theories designed around these “high impact” practices could be compromised. That is, demonstrable learning improvement requires clearly articulated SLOs that are tightly aligned with educational interventions and overarching program theory.

Moreover, although AAC&U provides VALUE rubrics for assessment purposes, it is unclear how these rubrics directly assess students’ skills or abilities. Research providing reliability and validity evidence for VALUE rubric scores is also lacking (Finley, 2011). As a result, many practitioners are left wondering how HIPs should be assessed and whether this assessment will be universal or streamlined across institutions. It is difficult to integrate assessment practice with learning improvement when the components of an educational intervention (e.g., high impact practices) exist in a vacuum – without explicitly aligned SLOs and assessment instruments.

AAC&U, among other national organizations, has contributed important pieces or components necessary to demonstrably improve student learning and shift higher education towards a learning paradigm (Barr & Tagg, 1995). They have provided VALUE rubrics, described theoretically important aspects of potent educational
interventions (e.g., HIPs), and cultivated inter-institutional connections – all of which are inarguably important standalone components.

However, the way rubrics and HIPs interface with one another or how they can be integrated to ensure learning improvement has not been well articulated or conceptualized. That is, akin to Sony and other technology corporations in the late 1990s, AAC&U is beginning to develop key components for linking assessment practice to demonstrable student learning improvement (e.g., assessment rubrics, HIPs, multi-state collaborations, etc.). The hope is that organizations like AAC&U will proceed as Apple did, with a focus on integration, rather than continuing to develop these components in isolation.

Building from the work of AAC&U and other organizations, practitioners finally possess the necessary components for learning improvement. But they still need to be shown how to fit them together in an integrated way (e.g., assessment practices integrated with educational interventions and program theory). I have a few suggestions.

Summary. Improvement is a goal of assessment. (Ewell, 2009); yet, assessment practice has too rarely contributed to demonstrable learning improvement – especially at the program level. I have explored potential causes of this issue, but what about a solution?

Fulcher and colleagues (2014) have postulated that higher education assessment practitioners and faculty members needed a more didactic, practical model: one that focuses on learning improvement; prioritizes program theory and educational interventions; and explicitly details how faculty can demonstrably improve learning. In addition, this model must focus not only on student-level interventions but also
intervening at the faculty level via appropriate development and training opportunities. Lastly, and most importantly, the model must explicitly integrate curriculum and pedagogy with assessment and learning improvement processes.

**Need for a Model to Integrate Assessment and Learning Improvement Processes**

Prior to dissemination of the Simple Model in 2014, higher education assessment practitioners did not have a model that explicitly integrated educational interventions and program theory with assessment and learning improvement. However, there were still isolated instances of national- and institutional-level efforts to enhance learning and promote student learning improvement (i.e., there were promising MP3 player prototypes, but no integrated music storage and management systems). In the following sections, I explore examples of such efforts.

**Examples of national-level learning improvement processes.** At the national level, initiatives such as the Wabash Study (Blaich & Wise, 2011), research related to Culture of Inquiry (Chaplot, Booth, & Johnstone, 2010), and the Lumina Foundation’s 2016 policies to improve higher education attainment have all attempted to promote alignment and improve student learning outcomes.

For example, researchers collaborated to define and describe a new “culture of inquiry” for higher education. As Chaplot, Booth, and Johnstone (2010) describe, the five stages in the Culture of Inquiry model, include: 1) defining a focus of inquiry, 2) gathering relevant and meaningful evidence, 3) engaging a broad range of practitioners in exploring the evidence, 4) translating collective insight into action, and 5) measuring the impact of action. Proponents of the Culture of Inquiry model stress that quantitative learning outcomes data alone are not enough; assessment results must be coupled with
appropriate insights from a diverse team of stakeholders. Under this model, student perspectives are prioritized as valuable and necessary data sources along with a broad range of practitioners (Chaplot, Booth, & Johnstone, 2010). Another key component of the Culture of Inquiry model is the creation of a “safe space” in which data can be openly explored and investigated with no fear of punitive repercussions (Chaplot, Booth, & Johnstone, 2010). Once data have been openly explored, stakeholders must collaboratively develop applicable action steps.

Another example comes from the Lumina Foundation (2016). Researchers describe strategies to transform higher education into a “learning system” in which students progress through intentionally designed learning experiences, gaining competencies along the way as they master various knowledge and/or skill domains. The model for creating “learning systems” includes three steps: 1) redesign curriculum around 21st century learning, 2) staff with well-prepared educational teams rather than individual course instructors, and 3) shift public policy narrative to reflect postsecondary education as a public good (Lumina Foundation, 2016). Further, pedagogy is delivered via teams of professors, administrators, and employers rather than faculty acting in isolation (Lumina Foundation, 2016). The Lumina Foundation’s model also calls for curricular re-designs that would allow students to “move through educational experiences to create personalized learning pathways,” as opposed to taking courses that fulfill program requirements in a checklist fashion (2016, p. 19).

**Examples of institutional-level learning improvement processes.** In addition to national-level initiatives to promote improvement, several isolated examples of learning improvement have arisen from specific institutions. At my own institution, James
Madison University (JMU), there are examples from general education and information literacy. Assessment specialists have assessed JMU’s general education programming for decades. Through rigorous methodology, including a university-wide mandatory assessment data and longitudinal data collection procedures, stakeholders at JMU realized that the general education curriculum required major revision. After the general education curriculum was modified, subsequent assessment data suggested that general education student learning outcomes had improved.

Similarly, JMU stakeholders annually assessed students’ information literacy skills through a required assessment embedded within general education communication courses. Information literacy assessment was also a graduation requirement for all undergraduate students. When outcomes assessment data suggested that students were not achieving the desired information literacy learning objectives, faculty from the libraries and educational technology departments re-vamped online information literacy learning modules. Ensuing re-assessment of student information literacy skills suggested learning improvements.

Other institutions have also evidenced isolated incidents of learning improvement. Researchers from the National Institute for Learning Outcomes Assessment (NILOA) conducted case studies of nine different institutions to investigate how institutions integrate assessment practice with student learning improvement. They concluded that promising practices include: time to reflect on assessment results, a clear vision for assessment practice, transparent communication, and involvement of key stakeholders like students (Baker, Jankowski, Provezis, & Kinzie, 2012).
Of the nine NILOA case studies, two institutions were able to bridge the gap between assessment and learning improvement: Capella University (CU) and Carnegie Mellon University (CMU) (Good, 2015). At CU, students in the psychology program were not achieving a particular learning outcomes. CU psychology students were unable to recognize ethical practice and the impact of diversity on ethical practice. Faculty used curriculum maps to thoroughly review alignment between learning objectives and educational interventions occurring within various psychology classes. They determined that there were few assignments and class objectives that actually addressed the given learning outcome. Then they worked with faculty and administrators to make informed modifications to the curriculum. For instance, they added modules and assignments “so that the outcome was addressed in a richer way” (Jankowski, 2011, p. 7). Follow-up assessment results suggested improved learning as a result of the modified curriculum (Jankowski, 2011).

At CMU, assessment processes were first incorporated into academic program reviews, which helped reinforce a culture of continuous improvement among faculty and staff. CMU also integrated their faculty development resources with their assessment practice by centralizing all assessment support within the Center for Teaching Excellence. Centralized infrastructure allowed faculty to more easily integrate assessment, pedagogy, curriculum, and program theory. It also facilitated faculty development opportunities needed to change educational interventions in a meaningful way. For example, assessment data revealed that engineering students were not meeting learning outcomes related to experimental knowledge. The Eberly teaching center helped faculty identify gaps in their curriculum and make appropriate modifications. The faculty
began teaching experimental knowledge in two new courses. Subsequent assessment results suggested that students’ skills improved as a result of the curricular changes (Kinzie, 2012).

Clearly, some initiatives and institutions have started to bridge the gap between assessment and student learning; however, they represent the exception, not the rule. None of the aforementioned examples focused on all of the necessary ingredients for demonstrable learning improvement. That is, some focused on aligning faculty development with assessment practices, while others focused on examining program theory and alignment of interventions to learning outcomes, or effective course redesign, or the importance of openly reflecting on data in a “safe space,” etc. In most instances, all of those components (and more) must be simultaneously present in order to evidence learning improvement.

As described in Chapter One, in 2014, Fulcher and colleagues disseminated a more explicit model for student learning that fully integrates assessment, pedagogy, curriculum, and program theory. Pervious researchers have certainly presented ideas about re-assessment, faculty development, and implementation fidelity. But Fulcher et al. (2014) fleshed out these ideas in a more intentional way. They provided a model and a detailed example that faculty could follow to demonstrably evidence improvements. The following describes how the model was successfully piloted with one academic degree program at JMU.

**The Simple Model in Action**

For the past two years, assessment practitioners and faculty development experts at JMU have been working with faculty from our Computer Information Systems (CIS)
program to pilot the Simple Model. First, CIS faculty used assessment data to select a learning area in need of improvement. They selected program requirements elicitation skills, which represent the ability of students to gather specific requirements of software programs. Elicitation skills require interacting with clients to understand their needs. CIS faculty openly discussed when and where these skills were taught. They realized that program elicitation skills were under-emphasized. That is, the program was not providing students adequate opportunities to learn.

To address this concern, CIS program faculty sought employers with requirements elicitation expertise. Employers helped describe requirements elicitation from the perspective of applied professionals. Then, CIS faculty worked with assessment specialists to design an analytic rubric that reflected the complexities of elicitation skills. The rubric ranging from 1-“Beginner” to 5-“Outstanding Experienced Professional.”

CIS faculty video-taped seniors demonstrating requirements elicitation skills before any pedagogical or curricular changes were implemented, and used the rubric to evaluate students’ proficiency. As anticipated, students’ elicitation skills were merely “Developing” (i.e., students scored “2” on average). To address this deficit, faculty used the videos and the rubric as educational aids. Specific video segments that demonstrated different levels of performance relative to the rubric criteria were identified. Faculty shared the video segments – with IRB approval – to the subsequent cohort of students. The purpose was to explicitly, visually demonstrate the elicitation process, delivered with varying degrees of quality. Use of the rubric as a pedagogical strategy was one among many used by CIS faculty as part of their new or modified intervention.
A year later, CIS faculty applied the same rubric to the next cohort of seniors, who had received the modified educational intervention compared to the previous cohort. Rubric ratings suggested that students’ elicitation skills had dramatically improved, after experiencing the modified intervention. Students’ rubric scores were, on average, 1.14 points higher than the previous cohort’s performance; from an average of 2 (Developing) to over 3 (Competent). On average, students who received the modified intervention scored a little over 3 standard deviation units higher on the rubric compared to students who did not receive the modified intervention. For reference, Cohen (1988) suggested that values of 0.2, 0.5, and 0.8 represented small, medium, and large effects, respectively. In this context, a 3.3 represents an enormous practical improvement in students’ elicitation skills.

Through work with the CIS program faculty, practitioners at JMU have advanced their understanding of how to successfully evidence improved student learning. Assessment, pedagogy, and curriculum were well-aligned. Despite the overall success of the CIS project, it lacked an important feature that should be included in learning improvement endeavors: implementation fidelity. While it is likely that the CIS faculty implemented their program with high fidelity – given the positive results – their causal inferences are limited because they did not collect fidelity data.

**Understanding Implementation Fidelity Research and its Importance to Learning Improvement**

Implementation fidelity is at the heart of scientific research. It has been applied in basic research design to promote internal validity. However, implementation fidelity was first assessed in practical or applied setting in the domain of health care. For instance,
researchers used implementation fidelity to study the delivery or implementation of psychotherapeutic techniques, as well as other psychiatric interventions and practices (Dhillon, Darrow, & Meyers, 2015). Although still rare, interest in educational applications of implementation fidelity research has grown steadily since the early 2000s (Dhillon, Darrow, & Meyers, 2015).

Implementation fidelity data are crucial aspects of curriculum review, modification, and validation. They allow us to move inside the black box noted in Figure 1, and better understand the (in)effectiveness of specific features of the given educational intervention (Cook & Shadish, 1986). Yet, assessment processes typically do not include empirical examination of the intervention via implementation fidelity research. Instead, practitioners incorrectly assume that the intervention as delivered is the exact same as the intervention as designed. Although practitioners may be able to implement a learning improvement model (e.g., the Simple Model), the validity of their inferences or conclusions will be compromised, unless the black box is illuminated.

In the following sections, I provide a hypothetical example to explain the concept of implementation fidelity research. The components of implementation fidelity and various data collection methods are also described. In addition, I discuss the importance of fidelity data for learning improvement.

**An implementation fidelity example.** Implementation fidelity research (i.e., the collection, analysis, and integration of fidelity data) can be conceptualized as a type of performance assessment. Johnson, Penny, and Gordon (2009) define performance assessment broadly as the act of demonstrating knowledge or skill by “engaging in a process or constructing a product” (p. 2). When researchers observe class sessions to
collect implementation fidelity data, they use a *fidelity checklist* that serves as a behaviorally-based performance assessment tool (Swain, Finney, & Gerstner, 2013). Consider the following hypothetical scenario of implementation fidelity research, based on the Computer Information Systems (CIS) example discussed previously:

Imagine that you worked with faculty from the CIS program. The faculty developed an educational intervention to help students better elicit program requirements for developing software systems. As part of the educational intervention, faculty articulated five skills students must demonstrate to successfully elicit program requirements.

Then faculty developed and agreed upon specific “intervention features” (e.g., activities, assignments, demonstrations, simulations, etc.) that they would implement. The intervention features were intended to help students achieve the five student learning outcomes. For one intervention feature, students complete an in-class activity during which they observed videos of fellow students practicing requirements elicitation. The students then discussed/evaluated the quality of their peers’ requirements elicitation skills via an analytic rubric.

The faculty worked with consultants to incorporate the aspects of this in-class activity into an implementation fidelity checklist. The checklist served as a behavioral-based tool used to assess the degree to which the *delivered* activity or intervention diverged from the *designed or intended* activity.

Four faculty members agreed to implement the aforementioned activity in their classes (e.g., have students view video and evaluate fellow students’ program elicitation skills). Two trained researchers attended class sessions, observed
faculty members, observed students, and completed the fidelity checklist based on their behavioral observations. That is, the two researchers observed faculty and student behaviors as they “engaged in processes” and “constructed products” related to the educational intervention or activity.

The fidelity data allowed researchers to answer questions such as “Did the faculty actually deliver or implement all designed features of the activity?”, and “To what degree were students responsive or engaged during the activity?” among other questions.

These hypothetical faculty clearly articulated their program theory via an implementation fidelity checklist. That is, they used the implementation fidelity checklist to align learning outcomes with specific “intervention features” (e.g., activities, assignments, demonstrations, simulations). The features were implemented to help students achieve the intended learning outcomes. Using the fidelity checklist, researchers were able to conduct an in-depth study of the intervention that faculty elaborated via the fidelity checklist. Thus, researchers could pair fidelity data with outcomes assessment data to distinguish the (in)effective features of the intervention.

Compared to the real-life CIS example discussed previously, the hypothetical version – using implementation fidelity data - would have enabled faculty to be more diagnostic. Faculty would have been able to make more refined, accurate judgments about which aspects of the intervention were effective. In the real life example – without fidelity data - CIS faculty could only conclude that students’ elicitation skills improved. They could not determine which features of the intervention were actually implemented,
how many times, with what quality, and so forth. Therefore, the CIS faculty could not determine which intervention features successfully contributed to learning improvements.

**Five components of implementation fidelity.** To further concretize the concept, consider the following five components of implementation fidelity: Program differentiation, adherence, time duration, quality, and responsiveness (Gerstner & Finney, 2013; O’Donnell, 2008).

(1) **What is the program differentiation for an educational intervention?** Faculty must clearly describe the intervention components and specific features (e.g., activities, demonstrations, discussions, assignments, etc.), which comprise the program differentiation component of implementation fidelity. Intervention components and specific features provide an operational definition of the educational intervention. They explicitly state the opportunities to learn and align them with specific student learning outcomes. Thus, all components and features can be considered theoretically essential to students actually achieving the learning outcome(s). The features are more detailed than the components. Together, the components and features should help students achieve specified student learning outcomes.

Program differentiation is informed by program theory. The underlying theory (e.g., cognitive, developmental, learning, etc.) informs the pedagogies and/or curricula that should be implemented to help students achieve a given outcome. The components and specific features that make up program differentiation provide a model for how a given intervention is supposed to work (theoretically). Recall, program theory is defined or comprised of specific curricula and pedagogy. Similarly, program differentiation is
comprised of components and specific features, which include educational activities (i.e., curricula) and instructional techniques (i.e., pedagogy).

(2) Was each of the specific features of the intervention actually delivered or adhered to? This is a simple “yes” or “no” question that fidelity researchers answer for each of the specific features. If the answer is “no,” then practitioners or faculty cannot make any conclusions or inferences regarding the effectiveness of that particular feature because the feature was not delivered. Additionally, if the specific feature was not delivered, yet students still achieved the learning outcome it was mapped to, then perhaps the specific feature is not essential to that outcome’s achievement. If a specific feature was not delivered or adhered to, students were not given the opportunity to learn material or skills that were considered, based on theory, essential for achieving the student learning outcome. When this occurs, it may be unfair to assess students on those learning outcomes (e.g., it is unfair to assess students regarding learning that is related to specific features that were not adhered to).

(3) Did each of the specific features last the intended amount of time? Faculty should decide approximately how much time should be devoted to each feature. Then fidelity researchers should record the amount of time actually spent. If the intended versus actual time allocation differs, student learning and subsequent performance on assessment instruments could be affected.

(4) To what degree was each specific feature implemented with quality? Ideally, faculty would implement all of the specific features with the utmost quality. If features are delivered in a disorganized, unclear, or confusing way, then students will not have an adequate opportunity to learn. When subsequent assessment results indicate less than
stellar learning, program faculty may have a better understanding of why (e.g., because fidelity data indicated that the material was not delivered with a high level of quality).

(5) To what degree were students responsive during the intervention? Imagine faculty adhered to all of the specific features, for the intended duration, and did so with the utmost quality. However, throughout the intervention, students were texting, talking, or sleeping. In other words, students were given opportunities to learn, but they did not take them. Thus, it is unlikely the intervention will have a positive influence on student learning. In this case, the intervention itself might be effective; however, because of students’ unresponsiveness, the true efficacy of the intervention is unknown.

All five components of implementation fidelity are captured and described on a fidelity checklist. Swain, Finney, and Gerstner (2013) provide a didactic explanation and an example of how to create a fidelity checklist; this description is beyond the scope of the current project.

Implementation fidelity general data collection methods. Across both K-12 and higher education settings, implementation fidelity data are collected in various ways. For instance, Tarr, Chavez, Reyes, and Reyes (2006) used teacher surveys and classroom observations to determine how textbooks influence teachers’ enacted interventions and what content sections of the textbooks are excluded from enacted interventions. Rowan, Camburn, and Correnti (2004) describe three general, fidelity data collection methods: Researchers observes classrooms, instructors complete standardized checklists or logs during or immediately after classes, instructors complete end-of-year questionnaires. Breitenstein, Gross, Garvey, Hill, Fogg, and Resnick (2010) also described video recording and audio recording as means to collect implementation fidelity data.
At JMU, assessment practitioners have mainly collected fidelity data via live class observations and/or video recordings using fidelity checklists (Swain, Finney, & Gerstner, 2013). Checklists provide a systematic way to capture all five components of implementation fidelity. Note, at JMU, fidelity data collection has been limited to student affairs programs.

**Importance of implementation fidelity research.** Implementation fidelity research is becoming increasingly important for researchers seeking external funding. For example, the U.S. Department of Education now requires grant recipients to measure and report implementation fidelity as an indication of educational program impact (Goodson, Price, & Darrow, 2015). In addition, public and private organizations are funding research to examine fidelity in educational contexts, develop best practices for fidelity research, and refine how fidelity is measured (Dhillon, Darrow, & Meyers, 2015; Hulleman & Cordray, 2009). Outside of grants and funding opportunities, faculty and educational researchers are also realizing the importance of fidelity data.

Implementation fidelity provides important information that can enhance the accuracy of the inferences made from outcomes assessment data (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001). For instance, without implementation fidelity data, one cannot know what actually happened inside the classroom (i.e., the black box). Practitioners do not know if the educational intervention they are trying to make inferences about was the intervention *as it was originally designed* or *some deviant of that*.

Fidelity data have important implications for assessment practice. If outcomes assessment data are unfavorable (e.g., students’ scores did not improve), fidelity data
provides necessary information to modify educational interventions. Often, when assessment results are unfavorable, stakeholders are left wondering why. With fidelity data, practitioners and faculty are better equipped to explain “why,” target specific areas that need modification, and make informed changes to the intervention. That is, perhaps assessment results were unfavorable because a feature was not actually adhered to, a feature was delivered with poor quality, and so forth (Dhillon, Darrow, & Meyers, 2015).

Alternatively, when outcomes assessment data are favorable (e.g., students’ scores improved), fidelity data can “provide a roadmap for replication” and help identify “critical ingredients of program success” (Bond, Evans, Salyers, Williams, & Kim, 2000, p. 79).

In summary, when coupled with outcomes assessment data, fidelity data indicate which features of educational interventions are (in)effective. Understanding the effectiveness of intervention features allows faculty to be more pedagogically efficient and intentional. They can avoid “wasting” time on features of an intervention that have been shown to be ineffective or unimportant for student learning improvement. In contrast, without fidelity data, it is difficult to determine whether unfavorable assessment results are due to a poorly designed intervention or incomplete/inadequate delivery of the designed intervention (Dhillon, Darrow, & Meyers, 2015). Lack of fidelity data can lead faculty to make one of two costly errors: abandoning interventions that are effective (but perhaps were not implemented with high fidelity), or continuing to implement ineffective interventions (Gerstner & Finney, 2013).

Engaging in implementation fidelity research can positively influence faculty buy-in to assessment and improvement initiatives. Recall, faculty work directly with
assessment experts to create the fidelity checklist. Therefore, their teaching styles and expertise are well-reflected and represented. Implementation fidelity research can also promote engagement with assessment processes. Every time faculty complete the checklist as part of a “self” audit, they contribute to the fidelity data collection process. Creating checklists and collecting fidelity data also help faculty achieve better alignment between their program theory, assessment, and intended student learning outcomes.

To create the checklist, faculty explicitly specify how student learning objectives align with specific intervention features, approximately how much time should be spent on various features, etc. The checklist also requires faculty to consider aspects of teaching that they may otherwise overlook, including the degree to which they delivered intervention features with high quality and student responsiveness during intervention delivery.

For these reasons, I conceptualize implementation fidelity research as its own kind of faculty development. It gives faculty the tools to *backward design* their courses (Fink, 2003), enhancing alignment between assessment, pedagogy, curriculum, and student learning. Fidelity data also help faculty synthesize assessment data, making results more meaningful and useful. Understanding how the *delivered* intervention differed from the *designed* intervention promotes more accurate conclusions, and informs changes to educational interventions (Fisher, Smith, Finney, & Pinder, 2014).

In general, literature guiding practitioners through the implementation fidelity research process, as applied to educational interventions, is lacking (O’Donnell, 2008). Perhaps this is why implementation fidelity has not been extensively applied in higher education contexts.
At JMU, assessment practitioners have successfully incorporated implementation fidelity research into student affairs programs including an alcohol early intervention program and a first-year orientation program. In 2012, the orientation program received the National Association of Student Personnel Administrators (NASPA) National Grand Gold Excellence Award for their implementation fidelity research. However, practitioners have not yet applied implementation fidelity research to a university-level, academic program. Before applying fidelity research to an academic program, it is useful to understand how fidelity research has been applied previously. Thus, I briefly analyze a sample of previous efforts to apply implementation fidelity research in K-12 and higher education contexts.

**Brief Analysis of Existing Implementation Fidelity Applications in Academic Contexts**

In the following sections, I analyze previous applications of implementation fidelity research and fidelity checklists. A sample of fidelity checklists and data collection methods are also examined. The analysis is conducted on K-12 applications followed by higher education applications. Although the current study concerns higher education, I included K-12 examples because they are more prevalent, and they may trickle up to higher education.

**Existing applications in K-12 contexts.** In K-12 settings, questions related to implementation fidelity data are often framed in terms of “opportunity to learn” and “educational equity.” That is, if students were not given opportunities to learn (e.g., the delivered intervention differed from the designed intervention), then it may be unethical, unfair, or inappropriate to test students on all aspects of the designed intervention.
Concerns about fairness and equity might help explain why there has been more application of implementation fidelity in K-12 education compared to higher education. In addition, federal legislation and policies like No Child Left Behind (NCLB) mandated high stakes assessment testing regimens and thus a profound need for fidelity data.

Kurz, Elliot, Wehby, and Smithson (2009) applied and studied implementation fidelity in eighth grade mathematics courses across 18 teachers. They were especially interested in comparing general (“regular”) classes and special education classes. The special education subpopulation of students might not have adequate opportunities to learn all of the content they are held accountable for knowing on large-scale assessment tests (Kurz et al., 2009). The delivered intervention might be differing from the designed intervention within special education classrooms, creating especially negative consequences for special needs students.

For their study, Kurz et al. (2009) distinguished three types of curricula or interventions: intended (e.g., what ought to be covered in the classroom), enacted (e.g., what teachers cover in the classroom and thus what students actually have the opportunity to learn), and assessed (e.g., what is measured on student achievement or assessment tests).

Kurz and colleagues required teachers to self-report the content topics covered and respective expectations for what students should know or do (e.g., akin to student learning outcomes) at three time points (i.e., beginning, midterm, and end of year). Teachers self-reported the level at which they covered 16 general content areas related to mathematics such as basic algebra, probability, etc. Student achievement of mathematics concepts was assessed via three different math achievement tests.
According to Kurz and colleagues (2009), alignment between intended and enacted curriculum with state standards was low for both general and special education classes. Also, on average, alignment was positively related to achievement. Therefore, interventions implemented with greater fidelity (e.g., the enacted intervention adheres to the intended intervention) may be related to more favorable student learning outcomes.

**Examples of fidelity checklists.** K-12 researchers use a variety of implementation fidelity data collection methods (e.g., teacher logs, teacher end of year surveys, observations, etc.), including checklists. The following describes a subset of checklists applied to K-12 research. For each, I briefly analyze: a) the breadth or comprehensiveness of the checklist, and b) the extent to which the checklist includes all five components of implementation fidelity.

I also draw from Popham’s (1997) rubric construction recommendations to analyze the checklists. Note, Popham’s recommendations are intended for rubrics. Implementation fidelity checklists can be considered a type of performance assessment rubric; therefore, Popham’s criteria were generally or “loosely” applicable.

The first checklist was used to study language arts education in third grade classrooms. Rowan, Camburn, and Correnti (2004) had teachers complete *Language Arts Logs* every day after class. The Language Arts Log – really a checklist – covered an array of topics, including:

- the duration or amount of time spent on language arts instruction that day,
- specific instructional actions teachers took during class that day,
- a rationale for why language arts were not covered that day (given the teacher indicated that zero time was spent on language arts), and
• to what extent various material (e.g., word analysis, writing, spelling, grammar, etc.) was emphasized that day (i.e., a major focus, a minor focus, etc.)

Each of these topics was further elaborated into detailed sub areas. Teachers reported the extent to which their instruction focused on each sub area. For instance, a teacher could report that she addressed “identifying story structure” during the class, and that it was only “touched on briefly” (Rowan, Camburn, & Correnti, 2004). Teachers also reported what students did during class to demonstrate comprehension. For example, teachers could report that students “answered brief oral questions,” “worked on a concept or story map,” “discussed text with peers,” etc.

Rowan and colleagues (2004) checklist was detailed and covered a comprehensive amount of content. Although their checklist contained more than the recommended three to five evaluative criteria for performance based rubrics (Popham, 1997), it was still user-friendly and logically organized. Also, even though the checklist was very detailed, it seemed like it would inform teachers rather than overwhelm them. The checklist adequately summarized and captured the delivered educational intervention.

The checklist, however, omitted ratings of quality of intervention implementation and student responsiveness. Perhaps teachers implemented all of the features of the designed intervention, but they did so with poor or low quality. Or maybe students were disengaged during the majority of the class activities, demonstrations, etc. Quality and responsiveness constitute important fidelity information that would not necessarily be captured by Rowan and colleagues’ checklist.
The second checklist was used to study a middle school life skills training program in California. More specifically, the California Healthy Kids (CHK) resource checklist was created to study a health-oriented, drug prevention program for adolescents. The checklist included adherence and timing or duration (i.e., number, length, and frequency of lessons implemented; each lesson should last ~45 minutes). Teachers could complete the checklist via self-reporting; alternatively, classroom observers could complete the log (California Healthy Kids Resource Center, 2007).

The CHK resources checklist was well-organized; it was streamlined and easy to interpret. Each section contained a reasonable amount of evaluative criteria (e.g., approximately three to five) (Popham, 1997). Additionally, the checklist provided a list of “suggested documentation” that could be included to further evidence adherence.

Nonetheless, the checklist had several weaknesses. Two important components of implementation fidelity were excluded (i.e., student responsiveness and quality of feature implementation). Moreover, the specific features were not adequately elaborated or detailed. Based on Popham’s recommendations (1997), the CHK Resource checklist would be considered too general. It did not fully delineate the nuances of the skills training intervention program.

The last K-12 checklist was created and disseminated by the National Center on Response to Intervention (NCRTI) (Mellard, 2010). Stakeholders from the NCRTI recognized the importance of implementation fidelity data and teachers’ need for guidance regarding fidelity data collection. To address this need, the NCRTI provided a “template” or “example” checklist for teachers who were developing their own checklists. The aim was for teachers to create individual checklists that examine
“curriculum and intervention strategies” in their classrooms. The NCRTI checklist was by far the most comprehensive of those reviewed. It demonstrated an impressive amount of breadth.

The NCRTI checklist also included an extensive users guide, including instructions for completing the checklist. However, the shear amount of information provided in the users guide might be overwhelming for teachers to digest (Popham, 1997). Assessment practitioners would likely need to provide supplemental training and/or support, ensuring that teachers could appropriately use the guide to inform their implementation fidelity practices.

In addition, the NCRTI checklist comprehensively addressed all five components of implementation fidelity. The checklist included features that were detailed (therefore adhering to Popham, 1997), yet flexible enough that it could be applied across various disciplines and classes. Moreover, the checklist features could significantly inform instructional design and guide teachers in developing lesson plans because they were adequately detailed.

Unlike other reviewed checklists, the NCRTI checklist included more detailed evaluations of adherence (note, the checklist included this in the “quality” section). For example, imagine the following hypothetical intervention feature: “The teacher makes the learning outcome(s) evident to the students during class.” In addition to providing a Likert rating scale (e.g., Yes, Sometimes, No, Unable to determine, etc.) to indicate adherence, the NCRTI checklist requires researchers to specify teacher actions that might be observed to support a rating of “Yes,” “Sometimes,” “No,” and so forth. Specifying specific teacher actions that would be indicative of each level of adherence (e.g., “Yes,”
“Sometimes,” “No”) could promote consistency in how the checklist is applied across different observers and classrooms (Tierney & Simon, 2004). Characteristics of the NCRTI checklist could potentially be adapted for use in higher education contexts.

Typically, trends in education “trickle up” from K-12 settings to higher education contexts (Blumenstyk, 2016; Trombley, 2001). As fidelity research in K-12 domains becomes increasingly effective and valued, implementation fidelity applications will likely burgeon throughout higher education. Currently, however, there are fewer examples in higher education academic settings compared to K-12.

**Existing applications in higher education contexts.** Australian researchers Bath, Smith, Stein, and Swann (2004) conducted a mixed method case study to determine the degree of alignment between a planned, an enacted, and an experienced general education intervention. The intervention focused on enhancing general, cross-curricular skills (i.e., communication, problem-solving, ethical and social sensitivity), in addition to discipline-specific knowledge (i.e., a Music program). The researchers conceptualized the three types of intervention (i.e., planned, enacted, experienced) similarly to Kurz et al. (2009) (i.e., intended, enacted, assessed).

Bath and colleagues’ (2004) study employed a qualitative variant of an implementation fidelity checklist. That is, they had faculty from a Music program articulate a designed intervention or curriculum, then they had students report their experience of the delivered curriculum. Specifically, at the beginning of the semester, faculty described how their course offered students opportunities to develop the cross-curricular skills (e.g., communication, problem-solving, etc.). At the end of the semester,
students completed a student experience survey indicating the extent to which the classes they completed actually helped them develop the aforementioned cross-curricular skills.

To better understand the student experience component, students and professors responded to open ended questions such as: “Which of these outcomes do you feel has had the greatest amount of time spent on it in this course?” and “Are there any outcomes included above that were not covered in this course?” Student and faculty data were compared to determine whether their perceptions of the implemented curricular interventions aligned. Bath and colleagues (2004) were also able to identify instances where faculty had intended students to experience certain intervention features, but students reported experiencing something other than the intended features.

By qualitatively comparing the two strands of data, researchers helped faculty identify and explore areas of misalignment or “low” fidelity. Faculty addressed gaps between their intervention intentions and students’ intervention experiences. Then, as an entire Music program, faculty made informed curricular modifications. Two years later, the Music program assessed students’ self-reported outcomes related to the cross-curricular skills (e.g., communication, problem-solving, etc.). They found that self-reported outcomes data were more favorable for students who experienced the modified curricula compared to students who did not. Results suggested that collecting a qualitative variant of implementation fidelity data might have positively influenced students’ self-reported general skills (Bath et al., 2004).

This study is important to consider because researchers applied some of the general concepts of implementation fidelity to an academic degree program. They
considered not only the *designed* versus *delivered* intervention, but also the students’ self-reported *experience* of the intervention.

An apparent limitation of Bath and colleagues (2004) research was that they used indirect indicators of student performance. The researchers did not report data from direct assessment measures, nor did they explicitly integrate fidelity data with outcomes assessment data. Also, they did not conduct live classroom observations or record data using a fidelity checklist. However, fidelity checklists have been applied in higher education contexts. The following briefly describes examples of such application.

**Examples of fidelity checklists.** For each higher education fidelity checklist, I briefly analyze: a) the breadth or comprehensiveness of the checklist, and b) the extent to which the checklist includes all five components of implementation fidelity. As noted previously, I *generally* apply Popham’s (1997) recommendations to analyze the checklists.

The first higher education fidelity checklist was applied to study Research Based Instructional Strategies (RBIS) (e.g., service learning, case-based teaching, etc.). Approximately 390 faculty members from chemical engineering, computer engineering, and statistics programs completed a survey indicating the amount of time they spent on different RBIS activities in their classes (Borrego, Cutler, Prince, Henderson, & Froyd, 2013).

Borrego and colleagues (2013) defined one to four *critical components* for each of the 10 RBIS; this constituted the *specific features* of the checklist. If a faculty member reported spending class time on a certain number of the specific features (e.g., class activities) associated with a given RBIS, then they were considered to have implemented
the RBIS with “high” fidelity. Thus, Borrego and colleagues’ fidelity checklist only addressed two components of implementation fidelity: program differentiation and adherence.

The checklist failed to capture quality, duration, and student responsiveness; three important components of implementation fidelity. It was not considered to have adequate breadth or comprehensiveness. Note, although the survey asked faculty to indicate what percentage of time they spent on the critical components of the RBIS (i.e., the intervention specific features), it did not ask them to indicate the designed duration. Consequently, researchers were unable to determine the extent to which the delivered duration of intervention features deviated from the designed duration. Furthermore, Borrego and colleagues did not appropriately integrate fidelity data with outcomes assessment data.

Based on Popham’s (1997) rubric recommendations, the checklist was too general. Perhaps the survey format data collection method or concerns for academic freedom hindered the researchers from creating a more comprehensive checklist. Data collection issues, lack of faculty buy-in, and threats to academic freedom are potential barriers to applying implementation fidelity in higher education academic contexts.

In contrast, the checklist elaborated program differentiation well (e.g., it clearly articulated the critical components or intervention features of each of the 10 Research Based Instructional Strategies). For instance, specific intervention features were clearly articulated for all RBIS. The intervention features were flexible enough that they could be easily applied to various course content, from engineering to statistics. Examples of intervention features for the Inquiry Learning RBIS included: “Student explains their
reasoning to another student while solving a specific problem” and “Students participate in a class activity that requires reflection and interaction in pairs or in a group” (Borrego et al., 2013, p. 410).

The second higher education checklist was used in a study of student peer mediation and English communication skills. Mamo and Yigzaw (2015) identified student participants from a randomly selected English Communication Skills class at Debre-Berhan University. Students were assigned to serve as either mediators or mediatees based on their English speaking abilities. Mamo and Yigzaw used a fidelity checklist to assess the degree to which a delivered Mediated Learning Experience (MLE) diverged from a designed MLE.

The MLE checklist contained 22 dichotomous “items;” each represented one specific feature of the intervention. Mediators and the mediatees completed the fidelity checklist “items” at the end of every mediation intervention. Students answered either “yes” or “no” regarding the occurrence of various MLEs (i.e., adherence). Researchers examined the level of agreement between mediator and mediatee pairs concerning whether or not various MLEs (i.e., specific features) were implemented. The MLE checklist included a comprehensive amount of specific features.

Mamo and Yigzaw’s fidelity had several flaws. It only addressed the program differentiation and adherence components of implementation fidelity, while ignoring quality, duration, and student responsiveness. Similar to Borrego and colleagues’ (2013) checklist, Mamo and Yigsaw’s (2015) was too general according to Popham’s (1997) recommendations. Also, the checklist might not yield data that can inform faculty’s instructional plans or course design in a meaningful way as Popham (1997) suggested it
should. Given there were no indicators of quality and responsiveness, the checklist may provide only a limited amount of actionable information.

The last fidelity checklist is from a student affairs program at JMU. The Orientation Office has used this fidelity checklist to assess their transfer student orientation programming for multiple years. Note, leaders from the Orientation Office received a nationally recognized award for research related to this checklist.

Three to four trained implementation fidelity researchers attend the transfer orientation programming each year. The fidelity researchers engage in an undercover role-play as transfer students, which allows them to participate in the transfer orientation programming as if they were legitimate transfer students. While doing so, the trained fidelity researchers use the checklist to collect data in real-time. During orientation, program facilitators are unaware of the identities of the fidelity researchers that are participating in the transfer programming. However, afterwards, the fidelity researchers help program facilitators integrate fidelity data with outcomes assessment data, facilitating more appropriate inferences given students’ assessment scores.

The transfer orientation fidelity checklist is quite comprehensive; it includes all five components of implementation fidelity (see Appendix A). The checklist is neither too general nor unnecessarily specific for Popham (1997); it contains an appropriate amount of detail. Moreover, data collected via the checklist helped inform program modifications in a meaningful way, as Popham (1997) said it should.

For instance, fidelity data helped program stakeholders realize that they needed to incorporate more interactive, meta-cognitive checks into their programming. Thus, Orientation staff created a handout to help students perform a meta-cognitive check of
their resources knowledge (e.g., knowing which on-campus offices provide various goods and services to students). They also had program facilitators review the correct answers with students as soon as they completed it.

Orientation program stakeholders were able to use fidelity data to make intentional modifications to their educational intervention (e.g., creating a handout, making students do a meta-cognitive check of their knowledge of university resources by completing the handout, having program facilitators review the correct answers to the handout to provide timely feedback). These modifications had a positive influence on student learning.

That is, after facilitators implemented the modified version of the educational intervention (i.e., the programming that included the handout and an immediate review of the answers to the handout), students’ tended to perform better on the resource knowledge assessment instrument compared to baseline data (e.g., data collected before the modified intervention was implemented). The Orientation example demonstrates learning improvement, made possible by the use of implementation fidelity data. Given fidelity data helped this program evidence learning improvement, additional studies connecting implementation fidelity to learning improvement are needed.

**Need for More Implementation Fidelity Research**

Two themes emerged from this analysis of implementation fidelity applications. First, several checklists were flawed because they were too general and did not address all five components of implementation fidelity. Perhaps practitioners and faculty need more guidance or better examples of implementation fidelity applications in academic programs.
Second, more implementation fidelity research needs to be done within higher education contexts (Coburn, Hill, & Spillane, 2016). Calls for quality assurance and demonstration of value added will necessitate more applications of implementation fidelity research. Encouragingly, checklists from K-12 settings could potentially inform such research. In addition to collecting fidelity data, researchers must integrate these data with well-aligned outcomes assessment data (Fisher et al., 2014; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). Integration is the key to making more accurate, informed inferences based on assessment results (Fisher et al., 2014).

At JMU, practitioners have applied implementation fidelity methodologies to student affairs programs. Through this dissertation, in partnerships with campus leaders and faculty members, I applied implementation fidelity research to an academic program. These efforts should advance the field of implementation fidelity research, bridging the gap between assessment practices and demonstrable learning improvement, particularly in academic programs.

**Purpose of Current Project**

This research project applies implementation fidelity to an academic program under the guiding framework of the Simple Model for Learning Improvement (Fulcher et al., 2014). In doing so, I aimed to:

- support faculty to develop an implementation fidelity checklist which will help communicate a learning intervention related to a University-wide academic program that focuses on enhancing students’ ethical reasoning (ER) skills,
- apply implementation fidelity methodologies to determine the fidelity with which that ER intervention was implemented, and
- integrate outcomes assessment data and fidelity data to determine the effectiveness of the ER learning intervention, demonstrate learning improvement upon re-assessment (i.e., from pre-test to post-test), and bridge the gap between assessment practice and learning improvement.
Accomplishing these aims would assist JMU in demonstrating learning improvement within a university-level context. Furthermore, practitioners at JMU would demonstrate learning improvements is a skill area (i.e., ethical reasoning) that is important to faculty, employers, and other stakeholders (AAC&U, 2013).

Students were assessed at the beginning of the fall semester (i.e., Assess); this provides baseline data. Then students experienced a new ethical reasoning educational intervention that faculty members created and articulated via an implementation fidelity checklist (i.e., Intervene). Note, this was a versatile checklist that was applied across different disciplines. At the end of the semester, students were assessed again to determine whether their learning improved (i.e., Re-assess). Pre and post-test outcomes assessment data, once integrated with fidelity data, suggested the (in)effectiveness of the intervention to demonstrably improve students’ ethical reasoning skills.

In addition, through this research, I helped JMU practitioners apply implementation fidelity to an academic program for the first time in our institution’s history. I also provided development opportunities to participating faculty members by training them in implementation fidelity research, a skill they can use to enhance instruction and course design beyond this project.

For the current research, the substantive area of application was ethical reasoning (ER) learning outcomes and skills. ER was chosen because:

- many members of the faculty, student affairs programs, and the administration were eager to help students improve their ER skills,
- ER education was a cross-disciplinary, university-level academic endeavor,
- the Madison Collaborative had well-defined ER SLOs,
- ER assessment instruments had copious supporting validity evidence,
- several years of data had already been collected to measure students’ ER skills, and
• stakeholders were not satisfied with students’ baseline levels of ER abilities (i.e.,
students were not meeting university-level strategic plan benchmarks for ER
skills).

Thus, the Madison Collaborative: Ethical Reasoning in Action was ripe for a learning
improvement project incorporating the Simple Model and implementation fidelity.

Introduction to the Madison Collaborative: Ethical Reasoning in Action

Beginning in the summer of 2011, as part of our institution’s Quality
Enhancement Plan (QEP) for reaccreditation, JMU stakeholders and faculty created the
Madison Collaborative: Ethical Reasoning in Action. The goals of the Madison
Collaborative (MC) focused on enhancing students’ ethical reasoning (ER) skills by:

• Providing a framework for teaching ER skills (e.g., the eight Key Questions or
8KQ),
• Offering ER development and training experiences for faculty, and
• Using assessment to demonstrate improved ER skills.

More recently, the MC began helping faculty learn how to teach ER skills using
evidence-based pedagogies.

To enhance students’ ER skills, MC stakeholders re-defined ER as a process
consisting of open-ended inquiries focused on multiple ethical considerations. The
multiple ethical considerations are conceptualized or framed by Eight Key Questions
(8KQ):

• Fairness - How can I act equitably and balance legitimate interests?
• Outcomes - What achieves the best short- and long-term outcomes for me and
all others?
• Responsibilities - What duties and/or obligations apply?
• Character - What action best reflects who I am and the person I want to
become?
• Liberty - How does respect for freedom, personal autonomy, or consent
apply?
• Empathy - What would I do if I cared deeply about those involved?
• Authority - What do legitimate authorities (e.g. experts, law, my religion/god)
expect of me?
• Rights - What rights (e.g. innate, legal, social) apply?

Then MC stakeholders operationalized their definition of ER via five cognitive student learning outcomes (SLOs):

1. Students will be able to state, from memory, all Eight Key Questions.
2. When given a specific decision and rationale on an ethical issue or dilemma, students will correctly identify the KQ most consistent with the decision and rationale.
3. Given a specific scenario, students will identify appropriate considerations for each of the 8KQs.
4. For a specific ethical situation or dilemma, students will evaluate courses of action by applying (weighing and, if necessary, balancing) the considerations raised by KQs.
5. Students will apply SLO 4 to their own personal, professional, and civic ethical cases.

The five SLOs represent what students should know, think, or be able to do as a result of experiencing MC interventions. To provide students opportunities to learn the knowledge and skills necessary to achieve these learning outcomes, the MC created ER learning interventions.

**Madison Collaborative Learning Interventions and Assessments**

**It’s Complicated.** As of 2016, the Madison Collaborative has implemented one systematic campus-wide ER intervention: *It’s Complicated*. Student completed It’s Complicated during orientation week right before they begin fall classes. All entering first-year students are expected to attend this programming. It assists new students with the academic and social transition to JMU (Smith, Fulcher, & Pyburn, 2015). Faculty members, student affairs professionals, and graduate students who have completed MC training facilitate the It’s Complicated intervention.

As a primer for It’s Complicated, students are instructed to read a scenario and watch a video about an ethical dilemma before beginning their Orientation programming.
Students are also asked to respond to an online survey that prompts them to think about ethical reasoning skills. Then, the week before classes begin, students participate in It’s Complicated. During It’s Complicated, facilitators introduce students to the MC 8KQ framework, and ask them to analyze the ethical dilemma using the 8KQ; thus creating a common intellectual experience. It’s Complicated is considered the initial direct intervention of the MC. It’s Complicated communicates the importance of ER (SLO 6) and exposes students to the 8KQ; thereby helping them understand each question’s meaning (SLO 1), and allowing them to identify the question(s) most relevant to a given scenario: SLOs 2, 3, and 4 (Smith, Pyburn, & Ames, 2016; Smith, Fulcher, & Pyburn, 2015).

**The Madison Collaborative Interactive (MCI).** Recently, MC stakeholders have also piloted a novel ER intervention in a general education communications class. The Madison Collaborative Interactive (MCI), is an online module-based intervention akin to a “choose-your-own-adventure” or evolving storyline. The MCI has eight modules. Each new module introduces a new KQ into the decision-making process and asks the student to use at least the episode’s primary KQ as well as other KQ to evaluate his or her choice. That is, each week, students reason through various ethical dilemmas using the 8KQ framework. Students get to interact with the story each week by voting on decisions. The next week’s 8KQ episode is dictated by the vote.

**MC Assessments.** Given the MC currently only implements one systematic 75-minute intervention (i.e., It’s Complicated), additional ER interventions were needed to help students achieve the MC SLOs. But first, to determine the extent to which students
achieve the SLOs, MC stakeholders worked with assessment specialists to create instruments that were well-aligned.

To assess students’ knowledge and skills related to SLOs 4 and 5, MC stakeholders created the Ethical Reasoning-Writing Rubric essay performance assessment (ER-WR) (see Appendix B) and accompanying essay prompt (see Appendix C). In addition, MC stakeholders created a multiple-choice test, the Ethical Reasoning Identification Test (ERIT), to assess students’ abilities related to SLOs 2 and 3. More detailed information about the psychometric properties of these assessment instruments is provided in Chapter three.

Note, the main focus of this research was to improve students’ abilities to evaluate courses of action by applying (weighing and, if necessary, balancing) the considerations raised by KQs) (MC SLO 4) and to apply SLO 4 to their own personal, professional, and civic ethical cases (MC SLO 5). These skills or learning outcomes are considered “higher-order” because they are more difficult to attain (e.g., in terms of learning taxonomies like Bloom’s, 1956). SLO 4 and 5 were assessed by the ER-WR essay performance assessment. Correspondingly, faculty mainly focused on SLOs 4 and 5 when developing their ER intervention during the summer of 2016 (See Chapter Three for more detail). They also used the ER-WR performance rubric to create the ER learning intervention (i.e., the specific features of the MC implementation fidelity checklist shown in Appendix F).

However, in addition to the ER-WR, students also completed the ERIT multiple-choice test, which measures students’ “lower-order” abilities to correctly identify the KQ most consistent with the decision and rationale (SLO 2) and identify appropriate
considerations for each of the 8KQs (SLO 3). Although somewhat secondary to the ER-WR data, students’ ERIT data could evidence important gains in students’ “lower-order” ER skills and abilities.

For instance, perhaps the ER intervention that faculty designed during the summer of 2016 was implemented with high fidelity and students’ skills related to SLOs 2 and 3 demonstrated large gains from pre- to post-test (i.e., their ERIT total scores were statistically and practically significantly higher at post-test compared to pre-test). Yet, students’ skills related to SLOs 4 and 5 did not demonstrate such gains (i.e., their ER-WR total scores were not statistically and practically significantly higher at post- compared to pre-test). Given this situation, one could conclude that while the ER intervention did not have the intended effect on students’ abilities to apply the 8KQ (i.e., SLOs 4 & 5), it was able to improve students’ abilities to identify the most appropriate KQ and the considerations for each KQ (i.e., SLOs 2 & 3). That is, while the intervention did not have the expected effects on students’ higher-order ER skills, it was still able to positively influence students’ lower-order ER skills.

Alternatively, given faculty members did not focus on SLOs 2 and 3, or the ERIT, when creating their ER intervention during the summer of 2016, one may expect to see smaller improvements in students’ abilities to identity KQ and their considerations (i.e., SLOs 2 & 3). Therefore, if students’ skills related to SLOs 4 and 5 demonstrate greater improvements (i.e., their ER-WR total scores were statistically and practically significantly higher at post- compared to pre-test) compared to their skills related to SLOs 2 and 3, this can provide initial validity evidence for the ER intervention.
In either case, students’ ERIT scores could reveal important conclusions concerning the ER intervention detailed in the MC implementation fidelity checklist. Thus, students were asked to complete the ERIT in addition to the ER-WR. Gains in ER-WR scores were the primary emphasis of the current project, with ERIT scores to supplement, as described above.

**Madison Collaborative’s Impact, So Far**

Given the aforementioned MC SLOs and aligned assessment instruments, MC stakeholders have assessed their impact on students’ ER skills, so far. Based on three years of assessment data, It’s Complicated has contributed to *immediate* gains in students’ ER skills. For instance, assessment results from fall 2013 and spring 2014 suggest that these two cohorts of students who received the It’s Complicated intervention scored higher on the ERIT *and* the ER-WR compared to students who did not receive any ER interventions prior to completing the ERIT and/or the ER-WR (Smith, Fulcher, & Pyburn, 2015).

However, these gains in ER skills may not be sustained over time. Students tended to score statistically significantly higher when they were assessed as first-year students in fall 2013 than they did when they were assessed again as second-year students in spring 2015; the magnitude of this decline was moderate (Smith, Fulcher, & Pyburn, 2015). For the fall 2014 cohort of students assessed again during spring 2016, on average, ER-WR scores did not change (Smith, Pyburn, & Ames, 2016).

Overall, as a result of It’s Complicated, students’ abilities to apply the ER process (i.e., MC SLOs 4 & 5) are either stagnating or declining over time, on average – based on ER-WR essay scores. In other words, students appear to demonstrate an initial, moderate
increase in ER skills as a function of a 75-minute intervention at the very beginning of their first semester (i.e., It’s Complicated). They, however, either make no subsequent gains in their ER skills related to MC SLOs 4 & 5, or backslide during their first 1.5 years on campus.

Recall, the MC has piloted a course-embedded, online module-based intervention lasting approximately eight weeks (i.e., the Madison Collaborative Interactive or MCI). Ames, Smith, Sanchez, Pyle, Ball, and Hawk (2016) conducted a study examining the efficacy of the MCI to enhance students’ ER skills. Based on results from students in a general education communications class, Ames and colleagues found that students’ ER-WR scores tended to increase after completing the MCI.

Yet, these initial gains in ER skills were not retained when students were re-assessed approximately one year after they completed the MCI. Thus, even after experiencing a longer eight-week ER intervention program (i.e., the MCI is a longer intervention compared to the 75-minute It’s Complicated intervention), students’ gains in ER skills, on average, were not maintained. These results may be expected given higher order ER skills (i.e., MC SLOs 4 & 5) are difficult to teach and difficult to learn.

Indeed, one study suggested that ER interventions might be more influential if faculty engaged in development or training opportunities prior to implementing ER interventions. Good (2015) explored the effectiveness of an ER week-long faculty development institute related to course redesign. Four faculty participated in a faculty development institute where they integrated or “infused” the JMU ER framework into their course curricula. Students enrolled in these “ER infused” classes the subsequent fall semester (e.g., the treatment group) and received the ER instruction that faculty had
infused into their courses – in theory, given no fidelity data were gathered. Good (2015) found that the treatment group’s scores on the ER-WR were statistically significantly higher relative to a control group of students that did not receive ER instruction from the four faculty member study participants.

In this case, the control group was scoring about marginal (1) on the ER-WR rubric, whereas the treatment group scored midway between marginal (1) and good (2) on the ER-WR rubric. Good’s (2015) results suggest that ER training and development opportunities for faculty members can potentially have positive effects on students’ ER abilities. While encouraging that the treatment group exhibited gains, their level of ER ability was nonetheless below the “Good” level on the ER-WR rubric. To attain higher ER skill levels, ER interventions may need to be longer and more intensive. In addition, implementation fidelity data need to be gathered and analyzed.

Acknowledging room for improvement, the MC has made strides towards defining, teaching, and assessing ER skills since its inception in 2012. MC stakeholders have articulated a definition of ER that lends itself to instruction, coupled with assessment instruments that yield reliable and valid scores (see Chapter Three). MC stakeholders have also prioritized faculty development and training focused on their definition of ER, the 8KQ, and ER assessment.

To further progress toward demonstrably improving every undergraduate students’ ER ability related to SLOs 4 and 5, the MC is making additional efforts to better understand the features of effective ER interventions. Given that understanding, they can create and implement additional ER interventions that are longer in duration and can be incorporated into classes across various disciplines. MC stakeholders can also
provide faculty development opportunities that will equip them to adequately deliver effective ER interventions.

MC stakeholders have essentially completed steps 1a through 2a of the Simple Model (i.e., 1a. Faculty nucleus dedicated to the learning improvement initiative, 1b. Administrative support of the learning improvement initiative, 1c. Rigorous assessment methodology involving a longitudinal data collection design, and 2a. Identify one or two SLOs to focus on). This study will help stakeholders complete the remaining steps (i.e., 2b. Investigate the current educational interventions already in place regarding the targeted SLOs and propose reasons why students might not be achieving these SLOs, 2c. propose learning modifications or create a new educational intervention, and 2d. detail timetable for educational intervention to be implemented and assessments to take place). Specifically, MC stakeholders will help faculty examine the current ER intervention, create a new ER intervention, study intervention implementation, and re-assess to determine whether students’ upper-level ER skills (i.e., SLOs 4 and 5) improved. If done correctly, completing the remaining steps in the Simple Model should eventuate in improved student learning.

**Research Questions**

During the summer, I taught faculty about implementation fidelity and helped them create a fidelity checklist. The faculty members’ checklist represented an MC ethical reasoning intervention aligned with SLOs 4 (*For a specific ethical situation or dilemma, students will evaluate courses of action by applying (weighing and, if necessary, balancing) the considerations raised by KQs*) and 5 (*Students will apply SLO 4 to their own personal, professional, and civic ethical cases*). I collected pre and posttest
assessment data using the ER-WR and ERIT, as well as implementation fidelity data, to address the following research questions:

**RQ 1: Is the observed variance in ER-WR scores mostly due to differences in students’ ER abilities? Are ER-WR scores reliable?**

a. Based on Generalizability Theory analyses, to what facet or component can I contribute the majority of the variance in students’ ER-WR scores?

a. How much variance is due to rubric element?

b. Based on Generalizability Theory analyses, what are the relative and absolute reliability estimates for students’ ER-WR scores?

**RQ 2: Did students’ ER skills improve from the beginning to the end of the fall semester? Do students who participated in the current study demonstrate greater ER abilities compared to students who did not?**

a. Did students’ ER-WR scores improve from pre to post-test as a result of experiencing the developed ER intervention?

b. Did students’ ERIT scores improve from pre to post-test as a result of experiencing the developed ER intervention?

c. Did students who participated in the study achieve higher ER-WR scores after experiencing the ER intervention than a randomly selected group of first-year and/or second-year students who didn’t experience the intervention and were assessed during the university-wide assessment day in fall 2015 or spring 2016 (i.e., “assessment day comparison group”)?

d. Did students who participated in the study achieve higher ERIT scores after experiencing the ER intervention than a randomly selected group of
first-year and/or second-year students who didn’t experience the intervention and were assessed during the university-wide assessment day in fall 2015 or spring 2016 (i.e., “assessment day comparison group”)?

e. Did students who participated in the study achieve higher ER-WR scores after experiencing the ER intervention compared to a group of students who experienced an ER-infused course as part of Good’s (2015) study?

RQ 3: To what extent did the delivered ER intervention differ from the designed ER intervention?

a. Overall, was the planned intervention implemented with high fidelity?

b. Which specific features were implemented with high fidelity?

c. Which specific features were not implemented with high fidelity

RQ 4: What insights can implementation fidelity data provide to help stakeholders make more accurate inferences from outcomes assessment results? Are naturally occurring differences in implementation fidelity related to outcomes assessment results?

a. Did students in classes where the faculty member implemented more of the intervention specific features, did so multiple times, with higher quality, and higher student responsiveness tend to show larger gains in their ER-WR scores than students in classes where the faculty member implemented only a few of the intervention specific features, did so infrequently, with lower quality, and lower student responsiveness?

b. For classes that demonstrated the greatest gains in students’ ER skills (i.e., via the ER-WR), which specific features were implemented, how
frequently, with what degree of quality, and what degree of student responsiveness?

Based on the answers to the aforementioned research questions, I recommend modifications to the ER intervention in Chapter Five.
CHAPTER 3

Method

Chapter three describes the processes and methods used to bridge the gap between assessment practice and learning improvement. First, I discuss a faculty-level educational intervention that aligns with step 2c in the Simple Model (*propose learning modifications or create a new educational intervention*). I then describe the two Madison Collaborative (MC) direct assessment instruments used to collect pre- and post-test outcomes assessment data. Descriptions of study participants and data preparation for fidelity, ER-WR, ERIT, and SOS data are also provided. I describe data management and motivation filtering procedures, and present sample sizes for analyses. Note, this research received IRB approval as of August 29th, 2016 (IRB Proposal No. 17-0085).

Faculty-Level Educational Intervention

As discussed in Chapter One, faculty-level interventions are important for student learning improvement. For the current research, faculty development was a major priority. To provide development for the seven faculty member participants in this research project, I facilitated a week-long institute from August 15, 2016 through August 19, 2016.

One goal of the institute was to educate faculty about implementation fidelity research and the Simple Model for learning improvement. Faculty were also charged with articulating and agreeing upon the specific features of an ethical reasoning (ER) intervention, aligned with MC SLOs 4 and 5 and the ER-WR (see Appendix B), that they later implemented during the fall 2016 semester.
I provide a schedule of topics covered and activities for all five days of the institute in Appendix D. The training workshop had the following faculty learning objectives. As a result of participating in the summer 2016 training institute, faculty will (see Appendix E):

- Explain how assessment practice and teaching and learning are connected or related
- Identify and describe the steps in the Simple Model
- Identify the five components of implementation fidelity
- Explain the steps or process of collecting implementation fidelity data
- Articulate why implementation fidelity data is important for demonstrating student learning improvement
- Discuss and agree upon the specific features of an effective ER intervention aligned with MC SLOs 4 & 5
- Design an ER intervention based on the agreed upon features that aligns with MC SLOs 4 & 5 and that can be applied in various classes, and
- Create a general implementation fidelity checklist aligned with the ER intervention and MC SLOs 4 & 5.

The training workshop had two main deliverables: an ER intervention that all faculty agreed to implement within their respective classes, along with an accompanying implementation fidelity checklist. Both were intended to be well-aligned to MC SLOs 4 and 5, as well as the ER-WR performance assessment instrument. Both deliverables were met by the end of the institute.

**Co-creating an ER intervention and fidelity checklist.** Figure 3 displays the general processes that faculty participated in during the institute to create the ER intervention and accompanying fidelity checklist. At the beginning of the institute, I trained faculty members in basic implementation fidelity research practices and the steps of the Simple Model. Then, for the remainder of the week, I worked with faculty to co-create an ER intervention and fidelity checklist aligned with MC SLOs 4 and 5 and the ER-WR rubric.
After faculty better understood implementation fidelity research and the Simple Model, we began to construct an ER intervention that could reasonably move students to an average score of 4-Extraordinary on the ER-WR rubric, starting with O’Donnell’s (2008) program components. Note, faculty had complete autonomy in deciding which level of intervention they wanted to create and teach; they chose to create a 4-Extraordinary intervention. However, instead of choosing to build a 4-Extraordinary intervention, faculty could have decided that they wanted to teach or implement a 3-Excellent ER intervention. The key is that the ER intervention was designed with the ER-WR rubric measurement tool in mind, from the outset. Thus, the intervention that the faculty created should be tightly aligned with the ER-WR assessment.

Initially, faculty generated a list of program components or general things that they would integrate into their classes that could help students become level 4-Extraordinary ethical reasoners. Faculty then participated in a series of Think. Pair. Share. exercises to co-create the program components, comparing and contrasting each other’s ideas to eliminate redundancies and combine components, where appropriate. Next, faculty shared specific activities, assignments, demonstrations, case studies, or other learning opportunities they have done in their classes in the past, or plan to do in the future, to help students achieve 4-Extraordinary ER skills.

After each faculty shared their plans, we categorized all of the specific activities according to the program components. Then, we worked backwards to make the specific activities general enough to be applied across the different courses and disciplines of each faculty member.
For instance, “Case studies or dilemma discussions” was one of the program components. Several faculty shared specific assignments from their class that were categorized as “Case studies or dilemma discussions.” As a group, we took these very course- and discipline-specific case studies and dilemma discussions and pulled out any underlying commonalities or similarities. These common threads became the specific features of the checklist as articulated by O’Donnell (2008). Specific features, coupled with the program components, thus defined the program differentiation (O’Donnell, 2008) for the faculty members’ ER intervention (see Appendix F).

Once the program components and specific features were articulated, the group of faculty critically reviewed them. During the review, faculty clarified certain language in the components and specific features; identified any instances where language/ideas were too prescriptive, specific, or limiting; and identified instances where language/ideas could be further detailed or made more specific.

Following the process described previously (see Figure 3), the faculty members created a general implementation fidelity checklist that could be used across all of their classes to collect fidelity data related to MC SLOs 4 and 5 (see Appendix F). Ethical reasoning subject matter expert, Dr. Bill Hawk, and implementation fidelity research expert, Dr. Sara Finney, also participated in this process during the training institute. Both helped faculty members articulate and further clarify the key features of an ER intervention that they would all agree to implement.

Faculty and Student Participants

Recall, seven faculty members participated in the week-long summer institute to create the Madison Collaborative ER intervention fidelity checklist. Faculty participants
taught six different courses (i.e., one course was team taught by two faculty members) in a variety of disciplines including: justice studies, health studies, science, philosophy, and education. Faculty participants were provided a $2,000 stipend to participate in the research study, which included the weeklong summer institute. However, faculty were highly intrinsically motivated to participate in the study; all indicated their interest in participating before knowing that they would receive a stipend. Faculty participants had a passion for improving their students’ learning and teaching their students ER skills. Also, faculty participants possessed moderate to advanced knowledge about the MC and the 8KQ framework prior to their involvement in the study. Several of the faculty had rated student ER essays during previous years, completed MC introductory training sessions, facilitated MC “Food for Thought” pedagogy discussions, taught the 8KQ in previous classes, etc. None of the faculty were inexperienced instructors; all had been teaching at JMU for multiple years and all but two were tenured.

Student participants were their consenting students. All student participants were asked if they were 18 years old on the consent form. If students were not 18 years old, they were not able to consent and therefore unable to participate in the research study. Note, faculty participants offered an “alternative” assignment or activity by which students could earn participation points or credit for the class, if they were younger than 18 years of age. To promote student motivation, faculty offered extra credit points and/or other class points to students who participated in the study.

A total of 289 students were in enrolled in the faculty participants’ courses. Of those 289 total students, 264 attended the pre-test sessions in Lakeview Hall (i.e., 91% response rate) between August 29th, 2016 and September 6th, 2016. Of those 289 enrolled
student, 242 attended the post-test sessions in Lakeview Hall (i.e., 84% response rate) between December 5th, 2016 and December 14th, 2016.

A summary of the data management, motivation filtering processes, and effective sample sizes is provided in Table 1. Student demographic information is provided in Table 2. After applying the data management and motivation filtering processes detailed in this chapter, the effective matched (pre/post) sample size was \( N = 191 \) for the ER-WR essays and \( N = 206 \) for the ERIT.

Note, nine “double enrolled” students were simultaneously enrolled in two classes (i.e., nine students were enrolled in both the health ethics class and the health diseases classes). Thus, these nine students ended up completing assessments twice at pre-test. One of these nine students did not complete the effort subscale and thus was completely removed from the dataset during motivation filtering. And the remaining eight “double enrolled” students’ duplicate records were removed such that they only appeared once in the dataset.

Given cross-sectional comparisons included in RQ 2, it is important to examine the demographic comparability of various student groups (See Table 2). Student participants from the current study were comparable to students assessed during assessment days in fall 2015 or spring 2016. As shown in Table 2, the distribution of ethnicities for students included in the current study and the assessment day comparison group were similar. The overwhelming majority of students self-identified as Caucasian, with less than 10% of students representing any other single ethnicity group (See Table 2). About 85% of the students in the current study were female, whereas about 61% of the assessment day students were female. Students from the current study and students
from assessment day were similar in age (i.e., 18-20 years old). Note that students in the current study were slightly younger, on average, compared to students assessed during assessment day in spring 2016 (See Table 2).

Similar to the current study’s demographics, about 80-90% of Good’s (2015) students self-identified as Caucasian. Less than 10% of students self-identified as any other single ethnicity. About 72% of the students in Good’s (2015) study were female, which was slightly less than the percentage of female students in the current study. Good (2015) reported that her “Control” group consisted of only freshmen and sophomore students. But she did not indicate the average age of study participants. It is unclear whether all of Good’s (2015) students were comparable in age to students from this study.

Gender was the only notable demographic difference between the assessment day comparison groups, Good’s (2015) students, and students included in the current study. However, the groups were still appropriately comparable, despite difference in gender percentages.

Students were also comparable in ways beyond demographic information. All of the groups included students from the same institution. Also, the same assessment instruments and standardized data collection procedures were used for students in the assessment day groups, Good’s (2015) study, and the current study. Good’s (2015) students participated as part a course they were enrolled in. Similar to the current study, Good’s (2015) students participated for either course credit or extra credit.
Madison Collaborative Assessment Instruments

In addition to fidelity data, I collected learning outcomes assessment data via two MC assessment instruments: the Ethical Reasoning Writing Rubric (ER-WR), and the Ethical Reasoning Identification Test (ERIT). The ER-WR is a direct measure of students’ achievement of MC SLOs 4 and 5. The ERIT directly measures SLOs 2 and 3. The following description provides more detailed information about each instrument, including psychometric properties, based on previous research.

The ethical reasoning writing rubric (ER-WR). The ER-WR essay is a performance assessment instrument that includes an ER essay scoring rubric and a prompt (See Appendices B & C). The ER-WR was designed to address the upper level MC SLOs: 4 (For a specific ethical situation or dilemma, students will evaluate courses of action by applying – weighing and, if necessary, balancing – the considerations raised by KQs) and 5 (Students will apply SLO 4 to their own personal, professional, and civic ethical cases), and thus was the main focus of the current research project. The ER-WR essay prompt asks students to consider an ethical situation or dilemma from their own lives, provide the considerations or perspectives from which they analyzed the issue, and explain how they ultimately arrived at their decision or solution.

The rubric used to score ER-WR essays has undergone several revisions since it was first developed in fall 2012, based on feedback gathered from raters in the first essay rating session in summer 2013. Scores are assigned to five rubric elements on a five-point scale (0 = Insufficient, 1 = Marginal, 2 = Good, 3 = Excellent, and 4 = Extraordinary). All raters had previous experience using the ER-WR rubric, and they all completed a three-hour rater training on using the behaviorally anchored rubric prior to rating student
ER-WR essay responses (Smith, Pyburn, & Ames, 2016; Smith, Fulcher, & Pyburn, 2015).

*Reliability and validity evidence.* The MC has been studying the reliability of students’ ER-WR scores for the past three years. Reliability of students’ ER-WR scores has varied over the years. For instance, in fall 2012 and spring 2013 ER-WR scores demonstrated favorable inter-rater consistency (teams’ G coefficients ranged from .91 to .97, and Φ coefficients ranged from .83 to .97) (Bashkov, Smith, Fulcher, & Sanchez, 2014). However, in fall 2014 and spring 2015, ER-WR scores yielded less than favorable consistency (G coefficient = .66, and Φ coefficient = .59) (Smith, Fulcher, & Pyburn, 2015). For the ER-WR scores for first-year students assessed during fall 2015, the G coefficient and Φ coefficients were 0.75 and 0.70, respectively, indicating adequate inter-rater reliability. ER-WR scores for second-year students assessed during spring 2016, the G coefficient and Φ coefficient were 0.66 and 0.58, respectively which represents slightly less than adequate inter-rater reliability (Smith, Pyburn, & Ames, 2016).

The 2012-2013 academic year served as the baseline data collection year for the MC (i.e., during fall 2012 and spring 2013 no Madison Collaborative interventions were in place yet). That is, the MC first offered the It’s Complicated intervention starting in fall 2013. Thus, students who responded to the ER-WR essay writing prompt during fall 2012 or spring 2013 experienced no Madison Collaborative interventions, and were expected to possess a negligible amount of ethical reasoning skills, as measured by the ER-WR essay rubric. Alternatively, students who responded to the ER-WR essay writing prompt during fall 2013, or later, experienced the It’s Complicated intervention, and thus should possess at least minimal ethical reasoning skills.
Expectedly, the average essay scores for the 2013-2014, 2014-2015, and 2015-2016 academic years were greater than they were for the 2012-2013 baseline academic year (see Table 3). That is, the cohorts of students who experienced the It’s Complicated ER intervention tended to score higher, on average, on the ER-WR rubric compared to the cohort of students who did not experience any MC ER interventions at JMU. In addition, ER-WR essay scores for the 2015-2016 academic year were greater than scores from any of the previous years (Smith, Pyburn, & Ames, 2016).

Assessment results have provided initial known groups validity evidence for the ER-WR essay scores. Three cohorts of students who were expected to possess at least minimal ER skills earned statistically and practically significantly higher scores, on average, than a cohort of students who were expected to possess negligible ER skills (See Table 4). Note, students assessed in the fall 2015 and spring 2016 cohort scored significantly higher, on average, than all three of the preceding cohorts (Smith, Pyburn, & Ames, 2016).

**The ethical reasoning identification test (ERIT).** The ERIT was designed to assess two of the lower level MC SLOs:

- 2 (*When given a specific decision and rationale on an ethical issue or dilemma, students will correctly identify the KQ most consistent with the decision and rationale*), and
- 3 (*Given a specific scenario, students will identify appropriate considerations for each of the 8KQs*).

The ERIT contains 50 multiple-choice items, it can be administered to a large cohort of students, and it can be easily scored using Scantron forms. Each of the items presents
students with an ethical decision or scenario that is best addressed by asking one of the Eight Key Questions (8KQ). Students must select from eight possible response options (i.e., one response option for each of the 8KQ) to indicate the ethical consideration most consistent with the given decision or scenario. For example, the following item presents an ethical scenario aligned with the Authority KQ because it represents the expectations of a legitimate authority (e.g., an attorney):

*Tommy finally reported his boss for sexual harassment to the company’s attorney.*

*The attorney said to Tommy, in no uncertain terms, ‘‘You WILL keep quiet about this incident.’’* Tommy complied and kept his mouth shut.

**Reliability and validity evidence.** Initial internal and external validity evidence for ERIT scores is provided by Smith, Fulcher, and Sanchez (2015). Reliability for ERIT scores has been shown to be adequate (i.e., >.70) across multiple years of assessment data (Cronbach’s α = 0.80 in fall 2013, 0.86 in spring 2014, 0.80 in fall 2014, 0.86 in spring 2015, 0.83 in fall 2015, and 0.86 in spring 2016).

Recall, students assessed in fall 2012 represent our baseline sample because they did not experience any MC ER interventions prior to completing the ERIT. The cohorts of students who completed the ERIT in either fall 2013, fall 2014, or spring 2015 (e.g., student who experienced the It’s Complicated intervention) all scored statistically significantly higher than the cohort of students from the baseline sample assessed during fall 2012, on average. The findings represent initial known groups validity evidence for ERIT scores.

Smith, Fulcher, and Sanchez (2015) provide further convergent and divergent validity evidence for ERIT scores, in addition to confirmatory factor analysis results.
supporting an essentially unidimensional factor structure underlying students’ responses to the ERIT items.

**Data Collection**

The implementation fidelity checklist developed during the summer institute was used throughout the fall 2016 semester to collect fidelity data from all seven faculty members’ classes. Faculty participants signed an informed consent form to grant consent for the researchers to observe their classrooms. The fidelity checklist (see Appendix F) was converted into an excel worksheet; thus, fidelity data were gathered and stored electronically. Having the fidelity data in an electronic format, as opposed to paper pencil, facilitated the process of adjudicating, averaging, and integrating the fidelity data with the outcomes assessment data (e.g., from the ER-WR and ERIT).

Each faculty member filled out the checklist for him or herself, as a self-report indication of fidelity (“self-audit”), for at least three class sessions throughout the semester. Faculty filled out “self-audit” checklists in different ways, depending on what was most feasible. For example, some instructors filled out the checklist for themselves for class sessions when the trained implementation fidelity researchers were not in class to collect data (e.g., providing three additional instances of fidelity data). Other instructors filled out the checklist for themselves on three days when the fidelity researchers were in class (e.g., providing reliability checks for existing fidelity data).

During fall 2016, I trained a group of seven graduate and undergraduate students to collect implementation fidelity data. This training was part of a three credit psychology research elective course (i.e., PSYC 403; class number: 73581) that met weekly. Students read numerous implementation fidelity research articles, which we discussed in class. We
also analyzed and reviewed example checklists. We reviewed the Madison Collaborative checklist to ensure that students were comfortable using the checklist to collect fidelity data. Students were also required to practice creating their own fidelity checklists, as well as explain the five components of implementation fidelity to the class.

The students and I observed a class together to practice collecting fidelity data. Afterwards, we debriefed about what aspects of fidelity data collection were most difficult and which features of the MC checklist required further clarification. Students discussed examples of what “low” quality of implementation may look like compared to “high” quality, among other topics. During several class sessions, we adjudicated fidelity data that students had previously collected. Periodic group adjudications promoted consistent data collection methods. Adjudications also allowed us to create “rules” for handling various data collection issues—in real time. For example, some of the checklist features were redundant with one another. Therefore, sometimes it was difficult to record the approximate duration for which specific features were implemented. As a group, we created a rule that we would all apply to promote consistency in our fidelity data collection: for a feature that was redundant with another feature, we would just enter “already counted” as the duration time; thus, duration would not be “double-counted”.

The aforementioned class experiences helped students increase their familiarity with fidelity data collection methods. Engaging in group adjudications and practicing data collection likely enhanced the consistency of fidelity data. Moreover, students did not know the faculty participants; students had never been enrolled in any of the classes they collected data from. Therefore, students’ ratings were unaffected by familiarity with the faculty members. All of the students’ fidelity data were shared with faculty
participants to review, which enhanced the validity of these data. The fidelity data collection training processes were important. Similar to providing validity evidence for the assessment scores, the previous information is given to support the veracity of the fidelity data.

The trained implementation fidelity researchers applied the checklist to six or more specified class sessions and/or specified class assignments throughout the semester. Ideally, two trained fidelity researchers would have attended faculty members’ designated class sessions. Then, these two researchers would discuss and adjudicate their fidelity data to ensure that one researcher did not overlook any specific features that were implemented, or that one researcher rated quality too low, etc. However, for any class sessions during which only one fidelity researcher was able to be present to collect fidelity data, the researcher followed up with the faculty member after class, sharing their completed fidelity checklist for that particular class session. They asked the faculty member to review the fidelity data they collected, making note of any specific features that were implemented that the fidelity researcher might have missed, commenting on whether or not the student responsiveness rating seemed accurate, etc.

In addition to classroom fidelity data, at the beginning and end of the fall 2016 semester, I collected outcomes assessment data from the students enrolled in the participating faculty members’ classes. Every student was asked to complete the ER-WR and ERIT Madison Collaborative assessment instruments once at pre-test (i.e., before their professors have implemented any of the features of the ER intervention created during the summer institute) and once again at post-test (i.e., after their professors have implemented the ER intervention created during the summer institute).
All pre-test data collection sessions were proctored by a trained graduate assistant student who had prior experience proctoring standardized assessment testing sessions; the same was true for all post-test data collection sessions. The proctor scripts used to administer the ER-WR and ERIT, during pre- and post-test data collection, precisely mirrored those used during the university-wide Assessment-day. When students arrived to the pre-test assessment testing sessions, they were asked to read and sign informed consent forms; they were asked to do the same when they arrived to the post-test assessment testing sessions.

During pre-test sessions, students responded to the ER-WR essay prompt (see Appendix C) using CARS Chromebooks and a software program created by CARS information security analyst, David Yang, M.S. (http://it-cars4.jmu.edu/APT/adayWriting.jsp). The secure program is used to administer all electronic writing assessments during JMU’s university-wide Assessment Day. The testing platform was created to reduce security threats that are inherent with other platforms (e.g., Qualtrics). Students had a maximum of 55 minutes to complete the ER-WR essay assessment. Students had to wait quietly for all other students in the testing session to finish the ER-WR essay before the proctor passed out the ERIT items, read the ERIT test instructions, and allowed them to begin the ERIT. On average, students took approximately 25 to 40 minutes to complete the ER-WR.

After completing the ER-WR, students responded to the ERIT via Scantron forms. For all testing session students completed the ER-WR essay assessment before they completed the ERIT. During the testing sessions, students did not have access to the ERIT items until the ER-WR assessment was complete. It is important to note this
because the ERIT provides students with all of the 8KQ; thus, it could bias their
responses to the ER-WR essay if they had just taken the ERIT and/or if they had the
ERIT items in front of them while completing the ER-WR.

To obtain an indication of students’ self-reported motivation during the data
collection sessions, students also completed the Student Opinion Survey (SOS) (see
Appendix G) after completing the MC assessment instruments. The SOS contains 10
Likert-type items and takes approximately 2-5 minutes to complete. It was administered
via Scantron forms right after the ERIT items. Students had a maximum of 45 minutes to
complete the ERIT and SOS items. Once they finished the ERIT and SOS items, and
handed in all of their testing materials, they were free to leave the data collection testing
session. On average, students took approximately 15 to 25 minutes to complete the ERIT
and SOS items.

Eighteen one-hour and 45-minute-long, pre-test data collection sessions were held
in Lakeview Hall during the fall 2016 semester beginning on August 29, 2016 and
concluding on September 6, 2016. Seventeen one-hour and 45-minute-long, post-test data
collection sessions were held in Lakeview Hall starting on December 5, 2016 and
concluding on December 14, 2016. Note, for sixteen of the eighteen pre-test data
collection sessions, all students completed both the ER-WR essay assessment and the
ERIT + SOS assessments.

For the remaining two pre-test data collection sessions, one faculty member
requested that their students be able to complete the ER-WR and ERIT assessments
during regularly scheduled class time (i.e., 9:30AM-10:45AM, and 11:00AM-12:15PM),
which only lasted one-hour and fifteen minutes. The faculty member cancelled her/his
class in lieu of the two assessment testing data collection sessions to further encourage students to come complete the assessment tests. Given the faculty member requested testing sessions that were shorter than one-hour and 45-minutes-long (e.g., they requested testing sessions that were one-hour and 15-minutes-long), their students only completed the ER-WR assessment and the SOS items (i.e., they were not administered the ERIT items).

To obtain pre-test ERIT data from this professor’s students, they were asked to visit the on-campus, Ashby Assessment and Testing Center to complete the ERIT. More specifically, these students were only permitted to visit the Ashby Assessment and Testing Center on a date after they had completed the ER-WR (i.e., post September 6th), but before their professor first implemented any features of the ER intervention in their classroom (i.e., October 3rd). In addition, the manager of the Ashby Assessment and Testing Center was given informed consent forms and instructions for the ERIT, which were given to students when they arrived for ERIT testing. For post-test, these students completed the ERIT after the ER-WR in Lakeview Hall, just like all of the other students. The Ashby Assessment and Testing Center was not used for post-test data collection.

Students who completed the ERIT at the Testing Center also completed the SOS, for a second time. That is, these students had already responded to the SOS once after they completed the ER-WR in Lakeview Hall during pre-test data collection. However, it was important to gauge students’ motivation a second time, given they completed the ERIT outside of the regular pre-test assessment testing session.
Implementation Fidelity Data Preparation

Prior to data analysis, I managed, cleaned, and integrated the implementation fidelity data. More specifically, I concatenated all implementation fidelity data within faculty member (i.e., within class). For three of the six classes, the faculty members taught two sections of each class, typically during back-to-back time slots. Given we were interested in making comparisons among faculty members (e.g., at the faculty level, not at the section level), I combined and averaged the fidelity data from across the two sections within the same faculty member. For example, if Faculty #1 received a quality rating of 4 for the specific feature “Review/Refresh 8 KQ” on the fidelity checklist for class section 1, but a quality rating of 3 for the “Review/Refresh 8 KQ” feature on the checklist for class section 2, then I recorded the quality for that particular feature as the average from the two class sections (i.e., 7/2=3.5), which aligned with averaging the performance of their students across these two sections.

For classes during which two researchers were able to attend and collect fidelity data, their adjudicated data were averaged. These adjudicated fidelity data were also shared with the faculty member to review. Some faculty participants also filled out the checklist for themselves for classes where two researchers were able to attend. In such instances, the faculty “self-audit” fidelity data were averaged with data from the two researchers.

For class sessions during which only one researcher was able to attend and collect fidelity data, their data did not need to be averaged with any other data. However, their fidelity data were reviewed by the faculty member for accuracy/completeness. For example, perhaps a faculty participant reviewed the fidelity data and noticed that the
researcher failed to capture one of the specific features that was implemented during class. The faculty member shared this feedback with the researchers. Then the fidelity data were adjusted to more accurately reflect all of the features that were actually implemented during that class session. Faculty participant reviews helped to promote consistent and accurate fidelity data. Some faculty participants also filled out the checklist for themselves for classes where only one researcher was able to attend. In such instances, the faculty “self-audit” fidelity data were averaged with data from the fidelity researcher.

For class sessions during which no fidelity researchers were able to attend, the faculty members filled out the fidelity checklist for themselves. The fidelity data for these specific instances could not be adjudicated because there was only one data source.

Fidelity data were stratified according to class/faculty member for comparison purposes. But stratified data are only reported in an anonymous way (i.e., Faculty #1, Faculty #2, etc.); therefore, individual faculty or classes cannot be identified or targeted. By stratifying fidelity data according to faculty/classes, I was able to examine differences in students’ ER abilities (as measured by the ERIT and ER-WR) that may be due to variability in the fidelity with which faculty delivered the ER interventions, within their respective classes (See Chapter Four).

**ER-WR Data Preparation: Rater Training and Essay Rating**

Prior to data analysis, the ER-WR essay responses were scored by human raters during a Madison Collaborative essay rating session held in Lakeview Hall on Tuesday, January 3rd and Wednesday, January 4th. The same process was used to score ER-WR data from the university-wide Assessment Day. The essays were rated by 11 veteran
faculty raters (i.e., raters who had previous training experiences with the MC and/or experiences using the ER-WR rubric) and one graduate student rater. Each rater participated in a three-hour rater training session before they read or evaluated the ER-WR essays. During the rater training session, the MC assessment liaison, Dr. Allison Ames explained each of the 8KQ, SLOs, ER-WR essay prompt (see Appendix C), and all elements of the ER-WR rubric (see Appendix B). Then, raters practiced applying the ER-WR rubric to three example training essays. As a group, they rated the first example essay one element at-a-time, allowing ample opportunities for raters to ask questions, explain the rationale for their ratings, receive clarification on how to appropriately apply the various rubric elements, etc. As raters evaluated example essays, Allison and I gauged whether they seemed to be appropriately calibrated to the rubric. Once the training portion of the session was completed, the raters independently rated their assigned ER-WR essays.

Prior to the ER-WR rater training and essay rating session, I removed the student ID numbers from all participants’ essays. The student ID numbers were replaced with a randomly generated seven-character ID number (e.g., 465999) as well as an alphabet letter (e.g., 465999H). The ID numbers had no meaning to essay raters; however, I used letter character in the essay ID numbers (e.g., “H”) to distinguish pre-test ER-WR responses from post-test ER-WR responses. I retained a file that linked the original student ID numbers to the “fake” ID numbers so that I could longitudinally match students’ ER-WR essay scores. Also, I assigned all raters into pairs; only I knew which raters were paired together. Raters were blind to partners to alleviate biases or influences. Thus, each student essay response was evaluated by two independent raters. Rater 1 and
Rater 2 evaluated and rated the same assigned subgroup of students’ ER-ER essay responses. Each rater partner evaluated and rated their assigned subgroup of student essays in reverse order to counteract fatigue effects. Each rater also rated one common “implant” essay during both days of rating. The implant essay was used to examine levels of rater “harshness.”

**ERIT Data Preparation**

Prior to data analysis, ERIT data were scanned and converted to electronic files. Note, ERIT data collected from Ashby Assessment and Testing Center were already in electronic form (i.e., downloaded from Qualtrics). ERIT items were dichotomously scored (i.e., correct/incorrect; 1/0) via SAS syntax; a missing response for an item was counted as incorrect. All 50 scored ERIT items were summed to create a total score for subsequent data analyses. Smith, Fulcher, and Sanchez (2015) provide statistical evidence to suggest that creating a total score with these data was appropriate.

**SOS Data Preparation**

Prior to data analysis, SOS data were scanned and converted to electronic files. Note, SOS data collected from Ashby Assessment and Testing Center were already in electronic form (i.e., downloaded from Qualtrics). Unlike the ERIT, the SOS was not scored dichotomously; instead, each item was scored using a five-point Likert scale. Further, certain SOS items were reverse coded (i.e., #3, #4, #7, and #9) (see Appendix G).

Recall, some students completed the SOS twice for the pre-test (e.g., once when they completed the ER-WR at the data collection assessment testing sessions in Lakeview Hall and once more when they went to the Ashby Assessment and Testing Center to
complete the ERIT). These students had two SOS scores for pre-test. If a student self-reported subpar motivation, their scores were excluded from analyses.

The SOS consists of two subscales: effort and importance (Sundre & Moore, 2002). Previous research, based on expectancy-value theory, suggests that the relationship between students’ self-report importance scores and their performance on cognitive assessment instruments is fully mediated by their self-reported effort scores (Cole, Bergin, & Whittaker, 2008). In other words, importance predicts effort and effort then predicts test performance. Furthermore, research suggests that students with self-reported effort subscale scores ≤ 13 should be deemed “unmotivated” (Rios, Liu, & Bridgeman, 2014; Wise & Kong, 2005).

I used students’ effort subscale scores as a motivation filter, rather than their importance scores or their total SOS scores. Students were deemed unmotivated if their effort subscale score was ≤ 13. Data from “unmotivated” students were removed prior to analyses. In the following paragraphs, I describe how SOS data were used to remove “unmotivated” students.

Before providing the results, I describe the data management and motivation filtering processes used to obtain the datasets used in analyses. To be included, students needed to have completed the ethical reasoning assessment at pre- and post-test; self-reported reasonable effort through the SOS motivation scale; and provided consent for their data to be used in this study. After describing these processes, I present the results in Chapter Four.
Data Management and Motivation Filtering Results

Table 1 provides a summary of the data management and motivation filtering process results. The following text provides details about those processes and how they affected the sample sizes used for subsequent analyses. Data management and motivation filtering for data collected during the pre-test are described first, for both the ER-WR and ERIT, followed by the same procedures applied to post-test.

Pre-test ER-WR. Unfortunately, the ER-WR data of four students were lost during pre-test. These students either failed to “save” their work before shutting down their assessment testing computers (i.e., Google Chromebooks), or students unknowingly being kicked off of the wireless network before they had “saved” their work. Thus, at pre-test, the effective sample size for the ER-WR essay assessment was 260 student responses out of the total number of enrolled students, 289 (i.e., 90% response rate).

Pre-test ERIT. A total of 172 students completed the ERIT during the testing sessions in Lakeview Hall and 80 students completed the ERIT in Ashby Assessment and Testing Center. Note, 86 of the 264 students who attended the assessment testing sessions in Lakeview Hall and completed the ER-WR were asked to complete the ERIT during a separate assessment testing session held in the Ashby Assessment and Testing Center per their professor’s request. Of those 86 students, 80 actually went to Ashby Assessment and Testing Center and completed the ERIT via a Qualtrics survey for pre-test. Thus, a total of 252 students (i.e., 172 in Lakeview Hall + 80 in Ashby Assessment and Testing Center) completed the ERIT at pre-test (i.e., 87% response rate).

Pre-test motivation filtering. Of the 252 students who completed the ER-WR and the ERIT at pre-test, four had self-reported effort subscale scores ≤ 13 and five did
not complete all of the items on the effort subscale; these nine students were removed from the dataset, bringing the effective sample size for the ER-WR and ERIT, at pre-test, to 243 students.

**Post-test ER-WR.** The ER-WR essay data for two students were lost due to students failing to save their work before shutting down their Chromebooks, or students unknowingly being kicked off of the wireless network before they had “saved” their work. Thus, at post-test, the initial sample size for the ER-WR essay assessment was 240 student responses (i.e., 83% response rate).

**Post-test ERIT.** A total of 242 students completed the ERIT during the post-test sessions in Lakeview Hall. Note, unlike pre-test, for the post-test no students were asked to complete the assessments in Ashby Testing Center; all students completed both assessments in Lakeview Hall.

**Post-test informed consent.** At post-test, two students completed both of the ER assessments but did *not* indicate on their student informed consent forms that they wished to participate in the study. Their pre- and post-test ER-WR and ERIT data were deleted and not used for any subsequent analyses. Deletion for non-consent brought the effective post-test ER-WR and ERIT sample sizes down to 238 and 240 students, respectively, who consented to participate in the study.

**Post-test motivation filtering.** A total of 236 students completed the ER-WR at *both* pre and post-test, and consented to participate in the research study. Of these 236 consenting students who completed the ER-WR at *both* pre- and post-test, 14 had self-reported effort subscale scores ≤ 13 and ten did not complete all of the items on the effort subscale. Therefore, these 24 students were removed from the dataset, bringing the
effective, matched sample size for the ER-WR to 212 students. The pre-test and post-test essays for these 212 students (i.e., 212 pre-test essays + 212 post-test essays) were independently evaluated by two trained raters as part of a Madison Collaborative essay rating session (described in Chapter three) held in Lakeview Hall on Tuesday, January 3rd and Wednesday, January 4th.

**“Unrateable” ER-WR essay data.** For a subset of the 424 ER-WR essays (i.e., 212 pre-test essays + 212 post-test essays) evaluated during the MC essay rating session in January, students did not exactly follow the ER-WR prompt instructions (See Appendix C). They did not actually write about an ethical situation, as they were instructed to do in the ER-WR prompt. Rather, they wrote about some event or occurrence that was framed as a difficult choice, they copied the “child stealing fruit” example given in the ER-WR prompt, etc. If students did not write about an ethical situation, then raters were allowed to deem their essay as “unrateable.” I reviewed all essay deemed unrateable to determine whether or not each was indeed unrateable.

Twenty-three of the total 424 ER-WR essays (i.e., 212 pre-test essays + 212 post-test essays) were deemed unrateable by two independent raters, reviewed by me, and subsequently verified as being “unrateable.” Fifteen of these essays were from pre-test and eight were from post-test. These 23 essays were not included in analyses. Furthermore, any corresponding pre or post-test essays were also removed prior to analysis. For example, if student A’s pre-test essay was deemed “unrateable” it was not included in analysis, and to preserve a completely matched pre-post sample size, student A’s post-test essay was also removed – even if raters had deemed student A’s post-test
essay to be rateable. This brought the matched pre/post sample size for data analysis for the ER-WR data to 382 student essays (191 pre-test essays + 191 post-test essays).

**Pre/post matched sample size for ER-WR and ERIT treated separately.**

Recall, 191 students completed the ER-WR at both pre- and post-test, consented to participate in the study, passed the effort subscore motivation filtering criteria, and submitted essays at both pre and post-test that were independently deemed “rateable” by MC essay raters (i.e., 66% response rate). Thus, the final pre/post matched sample size for the ER-WR assessment analyses was N = 191.

For the ERIT, 206 students completed the ERIT at both pre- and post-test, consented to participate in the study, and passed the effort subscore motivation filtering criteria (i.e., 71% response rate). Thus, the final pre/post matched sample size for the ERIT assessment analyses was N = 206.

Note, a total of 176 students completed both the ERIT and the ER-WR assessments at both pre- and post-test, consented to participate in the study, passed the effort subscore motivation filtering criteria, and submitted essays at both pre and post-test that were independently deemed “rateable” by MC essay raters. However, given the longitudinal and cross-sectional analyses were conducted separately for the ER-WR and ERIT assessments, I used the “matched within test” sample sizes because they were larger (i.e., ER-WR matched pre/post N = 191; ERIT matched pre/post N = 206; both ER-WR and ERIT matched pre/post N = 176).
CHAPTER 4

Results

Data Analysis & Results

Data analysis was conducted in January and February 2017. In the following sections, I describe analyses and results categorized by research question (RQ). Given faculty specifically developed the ER intervention to help students achieve SLOs 4 and 5, the majority of the data analyses focused on students’ ER-WR scores. The ER-WR instrument is intended to measure higher-order ER skills (i.e., SLOs 4 and 5). I treated the ERIT data analysis as supplemental because it was intended to measure students’ lower-level ER skills (i.e., SLOs 2 and 3). Yet, the ERIT data were still important given the implemented ER intervention may have positively influenced students’ lower-level ER abilities – (i.e., SLOs 2 and 3) in addition to their higher-order ER skills (i.e., SLOs 4 and 5).

Research question (RQ) one, while not the primary focus of the study, was important to address first as psychometric soundness of scores were critical to address other RQs. For instance, if ER-WR scores were found to be highly unreliable, subsequent analyses and interpretations of ER-WR data would be compromised. Indeed, RQ 1 could be conceptualized as merely an assumption that must be checked before addressing the substantive RQs (i.e., 2, 3, and 4). This study’s unique contributions and main thesis are described in RQs 2 through 4.

RQ 1: Is the observed variance in ER-WR scores mostly due to differences in students’ abilities? Are ER-WR scores reliable? ER-WR essay scores were analyzed using a statistical procedure called Generalizability Theory (G-theory), which is often
used to evaluate consistency in performance assessment scores. G-theory enables researchers to parse person variance – equivalent to CTT’s true score variance – from various error sources (i.e., facets). G-theory is also used to estimate relative and absolute reliability coefficients.

I used G-theory analysis with a multifaceted design (see Figure 4) to determine which facet (i.e., source of systematic error) contributed the majority of the variance in students’ ER-WR scores. The G-theory analysis had the following characteristics:

- Students (or Persons in G-theory vernacular) were the object of measurement,
- Raters constituted the first error facet, note that every rater did not rate every single students’ ER-WR essays; and
- Rubric Element constituted the second error facet, note that all raters rated student ER-WR responses on all five elements of the ER-WR rubric (see Appendix B).

Raters were considered a random facet because raters should be interchangeable, such that I would like to generalize to other raters outside of the specific groups of raters that actually reviewed students’ ER-WR responses. I treated rubric element as a fixed facet because I was not interested in generalizing to any other elements beyond the five elements that appear on the ER-WR rubric (See Appendix B).

Recall, each student’s ER-WR essay response was independently rated by two raters. And each rater pair was assigned a different subgroup of student essays to rate such that every single rater did not rate every single student response in the entire sample. As is common in large-scale assessment testing, it was not feasible for every rater to rate every student essay (DeMars, 2015).

For the essay data, there were two reasonable G-theory approaches that could be applied. First, raters could have been treated as being nested within student ER-WR
responses (i.e., (R:S)*E). Element, on the other hand, would be treated as crossed with raters and student response. The (R:S)*E) approach would represent a balanced, two-facet nested design (see Figure 4). Every rater used all five rubric elements to rate their assigned student essay responses. Therefore, all five levels of the rubric element occurred in combination with every level of rater and with every student response. Using this approach, a total of five variance components would be estimated. Variance due to:

- Student (Object of Measurement); $\sigma^2_S$
- ER-WR rubric element facet; $\sigma^2_E$
- Rater nested within Student; $\sigma^2_{R,SR}$
- Interaction of Student*Element; $\sigma^2_{SE}$
- Interaction of Student * Element * Rater interaction, plus random error; $\sigma^2_{ER, SER,e}$

Given the Rater facet was nested within the Object of Measurement, a main effect for Rater (i.e., systematic error due just to Rater) and the interaction effect of Student and Rater (i.e., Student*Rater) would not be estimated. In addition, nesting raters within students would not account for the fact that every rater evaluated essays from multiple students (DeMars, 2015).

Alternatively, I could have selected a different design that would enable estimation of variance due to Rater (i.e., a main effect for Rater). I could have used a crossed design, within rater pairs (i.e., (S*R*E) design conducted separately for each rater pair). To do so would have required separate G-theory analyses for each rater pair. Looking just within a single rater pair, the design would be fully crossed because within each rater pair both raters evaluate all student essays using all five rubric elements. In
other words, student (S) raters (R) and element (E) would be fully crossed with each other within each rater pair (P). Such a design would have allowed for the estimation of the main effect of Rater and the interaction effect of Student and Rater. Using this design, I would have conducted separate G-theory analyses for each rater pair and then taken the average across all rater pairs, weighted by the sample size within each rater pair, to calculate the G and $\phi$ coefficients (Chiu & Wolfe, 2002). A total of seven variance components would be estimated. Variance due to:

- Student (Object of Measurement); $\sigma^2_S$
- Rater facet; $\sigma^2_R$
- ER-WR Rubric Element facet; $\sigma^2_E$
- Interaction of Student with Rater; $\sigma^2_{S*R}$
- Interaction of Student with ER-WR Rubric Element; $\sigma^2_{S*E}$
- Interaction of Rater with ER-WR Rubric Element; $\sigma^2_{R*E}$
- Interaction of Student * Element * Rater interaction, plus random error; $\sigma^2_{ER, SER,e}$

Using the rater nested within student design (i.e., (R:S)*E), opposed to the fully crossed within pairs design, (S*R*E) within each pair (P), should introduce a negligible amount of bias into the variance components (DeMars, 2015). More specifically, when using the (R:S)*E design, the Student*Element variance component ($\sigma^2_{SE}$) will be negatively biased by a proportion of the variance due to the interaction between Rater and Element (Rater*Element). The Element variance component ($\sigma^2_E$) will be positively biased by a proportion of the variance due to the interaction between Rater and Element (Rater*Element). Lastly, the variance due to Student ($\sigma^2_S$) will be negatively biased by a
portion of the variance due to Rater ($\sigma^2_r$) (DeMars, 2015). As long as the total number of raters is large enough, bias in the variance components should be negligible (DeMars, 2015).

I chose to use the rater nested within student design, (R:S)*E, because it was the simpler and more practical alternative, as well as a common methodology used in large-scale assessment testing (DeMars, 2015). However, I also used the fully crossed within pairs design, (S*R*E), to examine low G and $\phi$ coefficients and restriction of range issues for pre-test essay scores.

I estimated the relative and absolute reliability estimates for students’ ER-WR scores. Relative reliability estimates (e.g., G coefficients) will always be larger than or equal to absolute reliability estimates (e.g., $\phi$ coefficients) because they reflect fewer error terms. For instance, the numerator of the equations to compute the G and $\phi$ coefficients are identical (i.e., variance due to Student or “true score” variance). However, the denominator of the equations differs because the relative error term used to calculate the G coefficient will always be smaller than the absolute error term used to calculate the $\phi$ coefficient. The relative error term consists of only those variance components that interact with the object of measurement (i.e., Student), whereas the absolute error terms consists of all of the variance components (expect the object of measurement), regardless of whether they interact with the object of measurement or not. Therefore, the relative error term for the (R:S)*E design consisted of only the variance due to the Student*Element interaction, the variance due to Raters nested within Students, and the error or “overlap” among the two facets and the Object of Measurement (See Figure 4).
Relative reliability estimates are most appropriate for making norm-referenced or “relative” decisions (e.g., comparing students to each other). Absolute reliability estimates are most appropriate for making criterion-referenced or pass/fail decisions (e.g., selection decisions, grouping students, when there is a standard or cut-score involved, etc.). The aim of the current research was to compare students’ ER-WR scores in a relative (e.g., how students score relative to one another, cross-sectionally and longitudinally) and absolute way (e.g., how students score relative to the scale or rubric criteria).

Ideally, the majority of the variance in students’ ER-WR scores should be attributed to variability due to student’s proficiency in ethical reasoning (e.g., “true score variance). That is, if the ER-WR assessment is sensitive to differences in students – and the error due to other factors is minimized – the variance component for Student ($\sigma^2_S$) should be the largest.

As shown in Table 5, for pre-test essays, the ER-WR rubric element facet ($\sigma^2_E$) contributed the most variance to students’ ER-WR scores. Much less variance was attributed to Student ($\sigma^2_S$) compared to the element ($\sigma^2_E$) and the rater nested within student ($\sigma^2_{R,SR}$) facets. This undesirable distribution of variability likely occurred because, at pre-test, the majority of the students scored similarly: they tended to earn low scores overall, but also tended to score higher on Element A and much lower on the other four rubric elements regardless of their scores on Element A. Moreover, ER-WR pre-test scores varied less compared to post-test scores; this difference was observed for average scores on each of the individual elements, as well as the overall average ER-WR scores.

The variance of the overall average ER-WR scores at post-test ($SD = 0.877, \sigma^2 = 0.769$)
tripled the variance of the overall average ER-WR scores at pre-test ($SD = 0.468, \sigma^2 = 0.219$).

For post-test essays, the majority of the variance in students’ ER-WR scores was attributed to the fact that different students’ with differing ER abilities produced the essays ($\sigma^2_S$) (e.g., “true score” variance) (See Table 5): a more desirable psychometric situation than was observed for pre-test essays. Less variance was due to the element ($\sigma^2_E$) and the rater nested within student ($\sigma^2_{R,SR}$) facets. Compared to pre-test, more variance was due to Student*Element ($\sigma^2_{SE}$). The interaction suggested that, at post-test, variance was attributed to the fact that some rubric elements tended to be easier (or harder) for some students to score well on than for other students. Students’ ER-WR scores were much more variable at post-test compared to pre-test, yet certain elements still appeared to be more difficult, on average, for students to score well on (e.g., Elements D and E).

Given the aim is to compare students’ ER-WR scores in a relative (e.g., how students score relative to one another, cross-sectionally and longitudinally) and absolute way (e.g., how students score relative to the ER-WR rubric criteria), it is important to estimate relative and absolute reliability estimates. As shown in Table 5, the $G$ (i.e., relative reliability estimate) and $\Phi$ (i.e., absolute reliability estimate) coefficients for pre-test were 0.340 and 0.269, respectively, indicating less than adequate inter-rater reliability ($< 0.7$). The relative/absolute error variances were 0.145 and 0.203, respectively. The $G$ and $\Phi$ coefficients for post-test were 0.841 and 0.814, respectively, which represents adequate inter-rater reliability ($> 0.7$). The relative/absolute error
variances were 0.122 and 0.148, respectively. Note, the relative and absolute standard errors were similar at pre-test (i.e., 0.381 and 0.451) and post-test (i.e., 0.349 and 0.385).

The low G and Φ coefficients for pre-test ER-WR scores were influenced by the restriction of range issue described previously. Recall, students tended to earn similarly low scores on the ER-WR at pre-test, and thus pre-test ER-WR scores varied little, especially compared to post-test ER-WR scores. Furthermore, the same group of raters evaluated all of the pre and post-test essays, during two concurrent days of rating. All raters evaluated a random mix of pre- and post-test essays (i.e., a rater might evaluate two pre-test essays, immediately followed by three post-test essays, followed by 1 pre-, then 2 posts, and so on). But raters were blinded as to whether they were evaluating a pre- or post-test essay. Thus, methodologically speaking, there should not have been large deviations in the consistency with which raters evaluated pre- and post-test essays. Given the reliability estimates for post-test essays were acceptable, perhaps the reliability estimates for pre-test essays would likely also have been acceptable – were there no restriction of range issues (i.e., had there been more variability in pre-test essay scores).

To further examine the potential effects of restriction of range on the estimated reliability of pre-test essays, separate G-theory analyses were conducted for each of the nine rater pairs who evaluated the pre-test essays (Chiu & Wolfe, 2002). Note, within rater pairs the G-theory design was fully crossed (i.e., both raters – within the same pair – rated all of the same assigned student ER-WR essays on all five of the rubric elements). The relative (G) and absolute (Φ) reliability estimates for pre-test, weighted by sample size within each pair and averaged across all nine rater pairs, were 0.505 and 0.348, respectively (See Table 6). Thus, the relative and absolute reliability estimates derived by
running nine separate G-theory analyses and calculating the weighted average across all nine rater pairs were slightly larger than those estimated from the \((R:S)*E\), rater nested within student, design (See Figure 4) (0.340 and 0.269, respectively) (See Table 5).

As shown in bold in Table 6, the relative and absolute reliability estimates for Pairs #6 and #8, indicated adequate reliability (>0.7). The relative reliability estimates for Pairs #1 and #9 also indicated adequate reliability. Table 6 also displays the amount of variance due to Student (i.e., object of measurement; variance due to the fact that different students with differing ER abilities composed the ER-WR essays). Note, the Pairs for which the G and/or \(\Phi\) coefficients were adequate, were the same Pairs for which the variance due to student (\(\sigma^2_s\)) was non-negligible. Given the raters in each of the nine Pairs were blinded as to who their rater partner was and all raters were trained during the exact same session using the same methods, one would not expect to see these differences in estimated reliabilities between rater pairs. Thus, the low reliability estimates at pre-test may be due to a restriction of range issue: too little variance due to Student because all students scored similarly poorly at pre-test.

In addition, as mentioned previously, the relative and absolute error variances for pre-test were similar to the error variances for post-test using the rater nested within student design (See Table 5). Post-test error variances were only slightly smaller than the pre-test error variances. When each rater team was analyzed separately, the relative and absolute standard errors were also similar across all nine of the Pairs, despite there being important between-pair differences in the estimated reliability coefficients (See Table 6). Similarities in error variances (or standard errors) suggest that the pre-test reliability estimates were artificially deflated due to a restriction of range issue.
Results from the fully-crossed G-theory analyses, coupled with similarities between error variances, suggested that low pre-test reliability estimates were not a problem. Restriction of range was the most likely cause of the low reliability estimates. Therefore, students’ pre-test essay scores were deemed reliable enough for use in subsequent analyses.

**Rater Harshness.** During both days of the MC rating session, all raters evaluated an implant or “common” essay. Raters scores on this implant essay were used to examine rater harshness (e.g., how stringent or lenient different raters tended to be). As shown in Table 7 and Table 8, each rater’s level of harshness tended to vary, somewhat, from the first to the second day of rating. That is, the same raters were not always the “harshest” or “easiest,” comparatively.

Moreover, these harshness data suggest that the raters were well-calibrated to the ER-WR essay rubric. On both Tuesday and Wednesday, the average essay score assigned by the harshest rater and the most lenient rater differed by only 0.7 and 0.8 points, respectively (on a scale of 0-4). Given the average scores assigned by the two most “extreme” raters (i.e., the harshest and the most lenient), on both days of rating differed by less than 1.0 point, evidence suggests raters were applying the ER-WR rubric elements consistently.

**Summary.** G-theory results for post-test ER-WR scores indicated adequate reliability estimates and the majority of variance in students’ ER-WR was due to Student (i.e., true score variance). The low estimated reliability coefficients for pre-test were acknowledged and further examined. Restriction of range was the likely cause. Thus, I
did not find definitive evidence to suggest that the pre-test ER-WR scores were so unreliable that they could not be used in subsequent analyses to address RQs 2, 3, and 4.

RQ 2: Did students’ ER skills improve from the beginning to the end of the fall semester? Do students who participated in the current study demonstrate greater ER abilities compared to students who did not? To determine whether students’ ER-WR scores improved from pre to post-test as a result of experiencing the developed ER intervention, I longitudinally compared students’ ER-WR total scores via a paired samples t-test. Similarly, I longitudinally compared students’ ERIT total scores. Note data met the proper assumptions (e.g., sampling distribution of the difference scores is normally distributed, etc.) for both analyses. Given some or all features of the piloted intervention were implemented with quality, students should demonstrate greater ER skills at the end of the semester compared to the beginning.

To determine whether students who participated in the study demonstrated greater ER abilities than students who did not, I conducted several cross-sectional comparisons using independent samples t-tests. For example, I compared participant and non-participant students’ assessment scores. Students assessed during fall 2015 or spring 2016 did not receive the ER intervention described in the MC Fidelity Checklist (i.e., “assessment day comparison group”). However, they did receive a brief 75-minute ER intervention as part of orientation programming (i.e., It’s Complicated).

I also compared participant students’ assessment scores to a group of non-participant students’ assessment scores from a highly comparable study. Students in Good’s (2015) study experienced an “ER-infused” class. Good’s (2015) students should have received more ER intervention than most students at JMU, who typically only
experience the *It’s Complicated* intervention. But fidelity data were not collected as part of Good’s (2015) study; therefore, it is unclear what ER learning intervention those students actually received.

In theory, participant students should achieve higher ER-WR scores compared to any of the aforementioned comparison groups. Participant students should also demonstrate larger magnitudes of improvement in their ER skills over time, based on the hypotheses presented by Good (2015). According to Good, students who experienced an intervention that was more tightly aligned with the ER-WR assessment should demonstrate ER improvements of larger magnitudes.

**ER-WR longitudinal.** On average, students who completed these courses improved their higher-order ER skills. Specifically, students’ overall post-test ER-WR scores (i.e., scores across all five ER-WR rubric elements) \( (M = 2.02, SD = 0.88) \) were statistically significantly higher than their pre-test ER-WR scores \( (M = 1.20, SD = 0.47) \), \( t(190) = 13.72, p < 0.0001, Cohen’s D = 1.74 \). On average, students’ post-test overall ER-WR scores were 1.74 standard deviation units higher than their pre-test scores. Such an improvement represents an effect more than twice as strong as what Cohen (1988) deemed large (i.e., 0.8). In addition, as shown in Table 9, students’ ER-WR scores for each of the individual rubric elements (i.e., A through E) statistically significantly increased from pre- to post-test. Therefore, these results suggest that the ER interventions students’ experienced during the semester positively influenced higher-order ER skills.

**ERIT longitudinal.** Recall, the ERIT was designed to measure students’ lower-order ER abilities, whereas the ER-WR was designed to address higher-level ER skills. On average, students who completed a course taught by the faculty participants improved
their lower-order ER skills from the beginning to the end of the fall semester. Specifically, students’ total post-test ERIT scores ($M = 36.13, SD = 7.14$) were statistically significantly higher than their pre-test ERIT scores ($M = 33.28, SD = 7.48$), $t(205) = 7.93, p < 0.0001, Cohen’s d = 0.38$. On average, students’ post-test ERIT scores were about four-tenths of one standard deviation unit higher than their pre-test overall ER-ER scores. At pre-test, students earned a score of 67% percent correct on the ERIT, whereas at post-test, they earned a score of 72% percent correct, on average. Both standardized and raw effect sizes suggest a small to moderate practical effect.

Given the ER interventions students experienced during the semester were aimed at the higher-order ER skills, it is not surprising that we observed a larger pre-post standardized effect for on the ER-WR (Cohen’s $d = 1.7$) than the ERIT (Cohen’s $d = 0.4$). One would expect to observe larger effects using an instrument more closely aligned with the intervention.

**ER-WR cross-sectional.** In theory, participating students should earn higher ER-WR scores than a randomly selected group of first-year and/or second-year students assessed during a university-wide Assessment day in fall 2015 or spring 2016 (i.e., “assessment day comparison group”). This hypothesis was supported. Participant students earned higher overall post-test ER-WR scores compared to students who did not participate.

More specifically, the average post-test ER-WR essay scores for participant students ($M = 2.02, SD = 0.88$) were statistically significantly higher than the average ER-WR scores for non-participating students who completed the ER-WR on a university-wide assessment day during fall 2015 ($M = 1.51, SD = 0.64$) or spring 2016 ($M = 1.21$,
Furthermore, as shown in Table 10, participant students earned statistically significantly higher ER-WR scores, at post-test, across all five of the rubric criteria compared to the two cohorts of non-participant students (i.e., students assessed in fall 2015 or spring 2016). Participant students scored approximately seven-tenths of one standard deviation unit higher or one standard deviation unit higher, respectively, on the ER-WR assessment compared to non-participant students assessed during fall 2015 or spring 2016.

Beyond statistical significance and the standardized effect size, students’ actual ER-WR scores matter. James Madison University’s strategic plan goal is for students to achieve an overall ER-WR score (i.e., average across all five rubric elements) of 2.0 by the year 2020. Results suggest that features of the piloted intervention can help students reach this university-level goal. It’s Complicated – the ethical reasoning intervention – currently deployed to all students has not helped students achieve a 2.0, which should not be surprising given that intervention is only 75 minutes long.

**ERIT cross-sectional.** In theory, participating students should achieve higher ERIT scores than a randomly selected group of first-year and/or second-year students assessed during a university-wide Assessment day. On average, participating students demonstrated slightly greater – yet not practically significant – ER abilities compared to students who did not. Participant students earned higher overall post-test ERIT scores compared to non-participant students (i.e., students who did not receive the piloted ER intervention). The average post-test ERIT score for participant students ($M = 36.13, SD = 7.14, N = 206$) was statistically significantly higher than ERIT scores for non-participant students assessed during fall 2015 ($M = 34.77, SD = 6.90, N = 465$), $t(669) = 2.330, p =$ \(SD = 0.57\). (See Table 10).
Similar to the results for the ERIT longitudinal comparisons discussed previously, the cross-sectional comparison represented a small practical effect. Although the piloted intervention may have a slight positive effect on students’ lower-level ER skills, differences between the cross-sectional comparison groups (i.e., student participants from fall 2016 v. non-participants from fall 2015 or spring 2016) were, on average, less than two-tenths of one standard deviation in magnitude.

**Across faculty comparisons.** One goal of this research was to examine how students’ ER abilities improved over time, and relate improvements to fidelity of intervention implementation. For instance, imagine that Faculty participant A’s students demonstrate the greatest improvements in their ER abilities. One might ask *Why? What did that faculty participant do pedagogically or what aspects of their implemented curriculum might have contributed to improvements in students’ ER skills?* Now, imagine that Faculty participant A implemented several of the features on the MC fidelity checklist (See Appendix F) with higher quality, with greater frequency, and higher student responsiveness compared to other faculty participants. If this were the case, results may suggest initial efficacy of the features included on the fidelity checklist. To answer this question, and thus study the efficacy of the intervention outlined in the checklist, students’ ER-WR scores were stratified and examined for each faculty member.

By examining stratified assessment results, I identify the class of students whose ER skills improved the most during the semester, to address RQ 2. Then, to address RQ 4
(What insights can implementation fidelity data provide to help stakeholders make more accurate inferences from outcomes assessment results? Are naturally occurring differences in implementation fidelity related to outcomes assessment results?), I integrate the “stratified by faculty member” results from RQ 2 with the fidelity results from RQ 3 (To what extent did the delivered ER intervention differ from the designed ER intervention?).

Recall, nine students were enrolled in two of the participating faculty members’ courses at the same time during the study (e.g., enrolled simultaneously in both Faculty Member A and Faculty Member B’s classes). However, those students are only counted once in the data sets; their data were not counted twice. That is, rather than having their data count for both faculty members, they were assigned to one Faculty member’s class (e.g., Faculty Member A). Therefore, the sample size for the other faculty member who they were not assigned to (e.g., Faculty Member B) was diminished or reduced by nine. In addition, recall that two faculty member participants team-taught one of the courses, meaning the study included seven faculty members implementing features of the intervention in six (not seven) different courses. For faculty data displayed in subsequent tables, the two faculty members who team-taught the course are only counted as one faculty member. Results are displayed for faculty #1, #2, #3, #4, #5, and #6.

**ER-WR.** As shown in Table 11, within all six courses, students’ higher-order ER skills showed (raw score) improvement for overall ER-WR scores and for all five rubric elements. More specifically, students’ overall ER-WR scores were statistically significantly higher at post-test compared to pre-test for five of the six courses (See Table 12). Both faculty #3 and #4 had a very small number of students included in the ER-WR
analyses. For faculty #3, although their students’ ER-WR scores were not statistically significantly higher at post-test, the change scores were practically significant. Students’ post-test ER-WR scores were nearly two standard deviation units higher than their pre-test ER-WR scores (See Table 12). Encouragingly, the median ER-WR score improvement, by course, was 1.7 standard deviations.

The students’ who completed the course taught by faculty member #5 demonstrated the largest improvements in their ER-WR scores from pre- to post-test, on average. These same students also demonstrated the highest overall ER-WR scores, on average, at post-test ($M = 2.983$, $SD = 0.523$) (See Table 11).

**ERIT.** As shown in Table 13, students’ lower-order ER skills showed (raw score) improvement from pre- to post-test. The students’ who completed the course taught by faculty member #4 demonstrated the largest improvements in their ERIT scores from pre- to post-test, on average (See Table 14). The students’ who completed the course taught by faculty member #2 demonstrated the highest overall ERIT scores, on average, at post-test ($M = 38.833$, $SD = 5.533$). Students’ ERIT scores improved about six-tenths of a standard deviation unit, on average, from pre to post-test.

Although (raw) average ERIT scores suggested improvements, students’ average ERIT scores were statistically significantly higher at post-test for only three of the six courses (at a more conservative alpha level of $.05/6 = .008$) (See Table 14). As discussed previously, both faculty #3 and #4 had a very small number of students included in the ERIT analyses, thus the significance tests may be underpowered. Across all faculty members, students’ gains in their lower-order ER skills were small to moderate.
**Results compared to similar study.** Assessment results from this study are comparable to Good’s study (2015), which used the same ER assessment instrument (i.e., the ER-WR) and a student sample from JMU collected during fall 2014. As discussed in Chapter Two, Good (2015) provided a course-redesign development opportunity for her faculty participants. During the course redesign, faculty created plans to “infuse” their courses with ER activities, lectures, etc. However, no fidelity data were collected during that study, so it is unclear what kinds of ER interventions the students in Good’s (2015) study actually experienced, how often they experienced them, with what quality they were delivered, and so forth.

Because I used the same assessment instrument and a similar student sample from the same institution, this study can be considered an expansion of Good’s (2015) study. Indeed, Good (2015) recommended this type of study. She hypothesized that implementing an intervention that was more tightly aligned with the ER-WR assessment instrument would produce even larger effect sizes. I did just that, and found precisely what Good (2015) hypothesized.

Students in Good’s (2015) study who experienced an “ER infused” course, on average, earned an overall ER-WR score of 1.47 at the end of the semester (See Table 15). Participant students, assessed during fall 2016, who experienced the piloted ER intervention ($M = 2.02, SD = 0.88$) scored statistically significantly higher than the students included in Good’s (2015) study who experienced the ER infused course ($M = 1.47, SD = 0.74$), $t(311) = 5.729, p < 0.001, Cohen’s d = 0.664$ (See Figure 5). On average, participant students scored about seven-tenths of one standard deviation unit larger than Good’s students.
Summary. Overall, students enrolled in all six courses tended to increase their ER skills from the beginning to the end of the semester. Students’ end of semester assessment scores also indicated that they demonstrated greater ER abilities, overall, compared to three different groups of non-participant students: students assessed in fall 2015 or spring 2016 who did not experience features of the piloted MC ER intervention described in Appendix F, and students assessed in Good’s 2015 study. Improvements in students’ higher-order ER abilities (as assessed by the ER-WR) were larger compared to improvements in students’ lower-order ER abilities (as assessed by the ERIT).

Importantly, participant students tended to reach (and in some cases exceed) the University strategic planning goal of being “Good” ethical reasoners (i.e., ER-WR score of 2.0) (See Appendix B).

Yet, there was “between-class” variability in how much students’ improved their ER skills. For instance, students’ who completed the course taught by Faculty #5 demonstrated greater improvements in their higher-order ER skills on average, compared to Faculty #1 (See Table 11). Given this variability between classes, perhaps students experienced different features of the MC intervention (See Appendix F).

Information concerning the fidelity with which the piloted intervention was implemented allow for more accurate interpretation of outcomes assessment results. For example, integrating assessment results with implementation fidelity data allows researchers to discern which features of the intervention the students taught by Faculty #5 received, with what level of quality, etc. The following section examines fidelity data to address RQ 3: the extent to which the ER intervention was implemented with high fidelity, across all six courses.
RQ 3: To what extent did the delivered ER intervention differ from the designed ER intervention? I concatenated all of the fidelity data within each faculty participant’s class to gauge:

- how frequently each faculty member implemented each of the specific features,
- the average quality with which each specific feature was implemented,
- the overall average quality with which all the specific features were implemented,
- the average student responsiveness for each specific feature,
- the overall average of student responsiveness for all specific features,
- the overall duration for the program components, and
- which specific features were implemented the most/least, on average.

Then I compared the fidelity data between all of the classes. Between class comparisons allowed me to determine if some faculty participants implemented the intervention with greater fidelity compared to other faculty.

When analyzing fidelity data, researchers typically examine four aspects: adherence, quality, student responsiveness, and duration. Adherence was captured as a simple “yes” or “no;” either the specific feature was implemented or it was not. Fidelity researchers recorded the quality with which faculty implemented intervention features, using a scale of 1 (Low- confusing) to 5 (high- clear). “High” quality of implementation can be characterized as faculty who clearly explained all instructions, answered students’ questions, provided clarification when needed, etc. Student responsiveness was based on
researchers’ perceptions of how engaged students were during intervention implementation. Similar to quality, perceived student responsiveness was captured using a scale of 1 (Low-unengaged) to 5 (High-engaged). An example of “high” student responsiveness would be students that were on task, actively participating in class, taking notes, discussing relevant topics when in small groups, answering questions, etc. Lastly, duration represented how long (in minutes) each feature was implemented. Ideally, actual duration is compared to intended duration for all intervention features. Imagine that a feature was intended to be implemented for 90 minutes, but in reality it was only implemented for 10 minutes. In this case, duration data would suggest that the feature was not implemented for enough time.

Unfortunately, faculty participants were unable to agree upon intended duration times for each specific feature on the fidelity checklist. They were also not comfortable articulating intended duration times for the more macro, checklist program components. Therefore, intended duration benchmarks were unavailable. To obtain duration data, fidelity researchers observed class sections and noted the approximate time that faculty members implemented specific features of the checklist. These duration times are rough estimates, too crude to be analyzed at the specific feature level. In addition, some of the features of the checklist tended to overlap with one another in practice (e.g., the “process something using the 8KQ” specific feature overlapped with several of the specific features on the “Case Study” program component). The overlap made it difficult to accurately and consistently record how much time was spent on each individual specific feature, without falsely inflating the amount of time spent on ER content (i.e., without “double counting” the time spent for certain specific features). Thus, duration data are
only presented and examined at the more aggregated program component level. These data should be interpreted somewhat cautiously, as rough estimates of time spent on each program component. Furthermore, future fidelity research studies can use these duration data to articulate intended duration times.

Fidelity data analyses allowed me to pinpoint specific features that were not implemented with high fidelity. Perhaps these features were too time consuming to implement, or they required a more in-depth, nuanced understanding of the 8KQ than these faculty members possessed, etc. I describe areas of low fidelity and propose hypotheses about why certain features might not have been implemented with high fidelity. I also solicited informal feedback from faculty member participants’ concerning why certain features may have been implemented with high/low fidelity throughout the semester. I include this information in Chapter Five, where necessary.

Note, fidelity was examined along a continuum; rather than as a dichotomous “all” or “none.” Fidelity is also discussed comparatively (e.g., a certain feature was implemented with “higher” quality in this class compared to other classes). If a given specific feature was implemented more frequently, with higher quality, and higher student responsiveness (comparatively) it may be considered a feature that was implemented with “high” fidelity, overall.

Given results for RQ 2, the ER intervention students experienced positively influenced their ER abilities, especially their upper-level ER skills. But the question remains: what specific features of the ER intervention did students actually experience? Implementation fidelity data answered this question and promoted more accurate interpretations of the assessment results presented for RQ 2. It was incorrect to assume
that all faculty members implemented all features of the intervention with equal
frequency, quality, etc. (e.g., that the designed intervention and delivered intervention did
not differ in any way). Therefore, without fidelity data, it would have been difficult to
understand precisely why students’ ER abilities improved.

Fidelity data also helped explain between-faculty member variability in students’
ER assessment scores. For instance, students who completed the course taught by Faculty
member #5 tended to score the highest on the ER-WR essay assessment at post-test. The
fidelity data demonstrated the specific features that Faculty #5 implemented, how
frequently, with what degree of quality, and how responsive her/his students were,
compared to the other faculty participants.

Note, information about each of the six classes is summarized in Table 16. I
provide a brief description of each course, the number of class sessions observed for
fidelity data, and number of assignments analyzed to collect fidelity data. Class
information is important to consider when integrating fidelity and assessment data. For
example, the six courses represent a range of disciplines, including health studies,
education, justice studies, science, and philosophy. The students enrolled in these courses
spanned the continuum of developmental statuses (i.e., freshmen through senior), as did
the course types. For example, one course fulfilled a requirement for general education
Cluster One, while others were electives taken by upper level students. Some courses
were primarily lecture based, while others were mainly seminar; several of the courses
included community service learning components, as well. The array of students and
course types was suitable for this particular study because the piloted ER intervention
was designed to be implemented across a variety of disciplines, course types, and student developmental stages.

In the following sections, I present aggregated fidelity data results (i.e., aggregated across faculty and aggregated up to the program component level of the fidelity checklist). Then I provide fidelity results, stratified according to faculty member, at the specific feature level of the checklist.

Results aggregated at program component level. Table 17 shows the fidelity results aggregated across faculty at the program component level of the fidelity checklist. Overall, the “Case Study” component had the highest approximate duration (in minutes). Faculty tended to spend the most time on specific features that fell within the “Case Study” program component. Student responsiveness ranged from an average of about four (mostly engaged) to five (engaged) across all six components of the checklist. Indeed, fidelity researchers noted that students were typically moderately to highly engaged during the classes they observed. Based on the adherence data, the largest number of features implemented were subsumed within the “Case Study” component of the checklist. Quality, like responsiveness, was rated fairly highly, ranging from an average of around three and half (moderate) to five (high) across the program components. Specific features subsumed with the “Examples” program component were implemented the least frequently and for the lowest duration of time, overall. The “Examples” program component received the lowest quality rating. The “Visualization” program component was also implemented infrequently compared to other components.

Across the program components, there was variability in the duration and adherence criteria. Duration ranged from less than five minutes (for the “Examples”
program component) to nearly two and half hours (for the “Case Study” program component), on average, across the program components. Across all classes, adherence ranged from six times for the “Examples” program component to 137 times for the “Case Studies” component.

Although each program component was implemented at least once, the variability of frequency is noteworthy. “Examples” and “Visualization” program components were implemented far less frequently and for shorter durations of time. Perhaps these features are less salient to the ER intervention. Alternatively, faculty may need additional training or development to help them understand how to implement the features subsumed within these two components.

Comparatively, the “Case Study” component might be important for influencing students’ ER skill development given it was implemented most frequently. Or perhaps the “Case Study” component is easier/more accessible for faculty to implement without additional training. Indeed, several faculty participants reported that they were already using case studies in their class before creating the MC intervention checklist. The aggregated fidelity data also suggested that students’ “Introduction/Foundational knowledge” and “Analysis” skills were not developed/practiced as thoroughly as their “Case Study” skills (See Table 17). If there were “deficits” in students’ foundational knowledge of the 8KQ and/or their abilities to analyze situations using the 8KQ, then they should be detectable in students’ ER-WR essay scores. RQ 4 will integrate fidelity data with assessment data to further examine this premise. Next, I present results at the specific feature level to explore how the delivered intervention differed from the designed
intervention. I also examine how fidelity data are related to improvements in students’ upper-level ER skills.

**Results aggregated at specific feature level.** Table 18 displays the fidelity data for student responsiveness, adherence, and quality stratified according to faculty member at the specific feature level of the fidelity checklist. Collapsing across faculty, there was great variability in the frequency with which each specific feature was implemented. For instance, one specific feature was not implemented by any faculty participants, whereas other features were implemented about 40 times (See Table 18). The “Students process something (debate, case, discussion, etc.) using 8KQ” specific feature was the most frequently implemented. However, given this feature tended to overlap with other checklist features, this frequency count might be inflated. The “Students identify where/how each of the 8KQ are/are not applied within the case” and “Students give/discuss rationale for how each of the 8KQ are/are not applied” specific features were also implemented with comparatively high frequency.

The “Identify and explain how characteristics or features make the case (in)effective” feature was not implemented at all. Given the feature was not implemented, we can only hypothesize regarding its relevance to the ER intervention. Nevertheless, it is curious why no faculty members implemented this feature. Faculty may require additional training on how to deliver this feature.

Student responsiveness ranged from three (somewhat engaged) to five (engaged), and quality ranged from two (low to moderate) to five (high). Looking within faculty members, student responsiveness ranged between four (mostly engaged) and five (engaged), and quality ranged between four and four and half (moderate to high), similar
to the aggregated program-level results. Faculty #5 tended to have the highest rated student responsiveness, whereas Faculty #2 tended to have the lowest; however, both received fairly high student responsiveness ratings (i.e., approximately 4 and 5, respectively). Faculty #1 received the lowest overall rating for quality, just barely below 4. Faculty #4 and #5 received the highest overall rating for quality, approximately 4.5. In terms of frequency or adherence, some faculty participants implemented double or even triple the amount of features compared to others. For example, Faculty #2 implemented the highest frequency of specific features. Faculty #1 implemented the lowest frequency of specific features. If these specific features are important aspects of an effective ER curriculum, then one might expect students from Faculty #2’s class to perform better on the ER-WR than students from Faculty #1’s class. As shown in Tables 11, 12, 13, and 14 this was the case.

The checklist contained a total of 22 specific features (See Appendix F). Faculty #1 implemented seven of those 22 features at least once (i.e., 32% of the total number of features). Both Faculty #2 and Faculty #4 implemented 16 of the 22 features at least once (i.e., 73% of the total number of features). Similarly, Faculty #3 implemented 18 (i.e., 82% of the total number of features), and Faculty #5 implemented 17 of the 22 features at least once (i.e., 77% of the total number of features). Faculty #6 implemented nearly all of the 22 features at least once (i.e., 21 features implemented; 95% of the total number of features). In theory, if all features on the checklist are truly salient to improving students’ ER abilities, then the students who completed courses taught by faculty members who implemented more of the 22 specific features should tend to be the same students who earned higher scores on the ER-WR (i.e., who demonstrated greater ER abilities).
Figure 6 displays the number of total intervention features implemented by each faculty member in relation to their students’ average overall post-test ER-WR essay scores. From the figure, it appears that the relationship between number of total specific features intervention implemented and students’ upper-level ER skills is positive, but not exceptionally strong. Notice that faculty member #6 implemented nearly all of the 22 features on the checklist at least once, yet her/his students did not demonstrate greater ER abilities, on average, compared to Faculty #4 and #5, who implemented fewer of the total 22 specific features. In addition, Faculty member #3 implemented nearly 3 times as many of the 22 total specific features, at least once, in their course compared to Faculty member #1. Yet, Faculty member #3’s students demonstrated similar (raw score) improvements in their upper-level ER abilities as Faculty member #1’s students (See Table 1). The relationship between frequency of adherence and students’ ER abilities may be complex; it may depend on additional aspects of fidelity, beyond just the frequency of feature implementation. Perhaps other aspects of fidelity, like quality and responsiveness, are more important than frequency of adherence. This premise was further explored to address RQ 4.

**Summary.** Fidelity results aggregated at the program component level and specific feature level revealed similar patterns. Overall, the “Case Study” component seemed to dominate the ER intervention implementation. That is, faculty tended to spend the most time and implement the most specific features subsumed within the “Case Study” component. Perhaps this component was thought to be the most important. Or, maybe it was “easier” to implement pedagogically compared to other components? Alternatively, maybe faculty found that their existing course content integrated/aligned
with the “Case Study” component more conveniently or efficiently than the other components?

Among faculty participants, there was some variability in student responsiveness and quality ratings, but both were typically rated moderately high (4) to high (5) across. However, duration and adherence differed between faculty members. As discussed, some faculty members implemented more of the 22 total specific features and implemented specific features with greater frequency, compared to others.

Encouragingly, all of the faculty members (except one) were able to implement most of the 22 specific features included on the MC checklist, at least once during the semester. Furthermore, the majority of the specific features were actually implemented multiple times throughout the semester (See Table 18). At the more macro, program component level, the interventions appeared to be implemented with fairly high fidelity. At the more detailed specific feature level, the intervention appeared to be implemented with greater fidelity for some faculty members compared to others (e.g., Faculty #6 compared to Faculty #1).

RQ 4 integrates fidelity results with assessment results (RQ 2) to support more appropriate inferences from students’ assessment scores. Additionally, RQ 4 explores the degree to which the between-faculty differences in implementation fidelity are related to improvements in students’ ER skills.

RQ 4: What insights can implementation fidelity data provide to help stakeholders make more accurate inferences from outcomes assessment results? Are naturally occurring differences in implementation fidelity related to outcomes assessment results? In general, the purpose of data integration is to understand which aspects of the
intervention were salient to improving students’ upper-level ER skills. Also, the data integration will demonstrate how intervention implementation was related to improvements in students’ ER skills. Implementation fidelity data were integrated with outcomes assessment data for three reasons:

- to help MC stakeholders and faculty make more accurate inferences concerning students’ ER-WR and ERIT scores,
- to understand the (in)effective features of the ER intervention, and
- to help explain why students’ ER skills improved.

To better understand how fidelity and assessment data were integrated, imagine students’ ER-WR scores demonstrated statistically and practically significant increases over time. Imagine that implementation fidelity data indicated that faculty consistently and frequently implemented the following specific features, from the fidelity checklist, throughout the semester:

- Elaborate or unpack each of the 8KQ (e.g., reviewing the handbook, lecturing, PowerPoint slides, video clip, discussion, etc.);
- Map 8KQ to some other work; and
- Give/discuss rationale for how each of the 8KQ are/are not applied

However, faculty were rarely able to implement any of the other specific features of the ER intervention. Uneven implementation may suggest that:

- the three aforementioned specific features are the most salient to an effective ER intervention, and/or
- the remaining features that were not frequently or consistently implemented might not be as important or salient as faculty initially thought.
Alternatively, imagine that the fidelity data had indicated that faculty implemented all of the specific features outlined in the checklist; however, the three aforementioned features were the only features implemented with high quality. In this case, perhaps the remaining specific features are still salient to an effective ER intervention. Nevertheless, the true effects of the features were not evidenced in the student learning outcomes assessment data because they were not implemented with a high level of quality.

Perhaps the implementation fidelity data will suggest that one of the specific features that faculty thought would be salient to an effective ER intervention is actually not. Imagine that faculty implemented a specific feature infrequently, with low quality, and low student responsiveness, but students’ ER skills still improved. Faculty members and MC stakeholders may decide not to include that feature as part of their ER interventions. Alternatively, stakeholders may retain the specific feature, even though it is not as salient to improving students’ ER abilities. Then, faculty would not need to heavily emphasize that particular feature.

It is expected that students who experienced classes where the faculty member:

- implemented all of the intervention specific features,
- did so multiple times,
- with high quality, and
- high student responsiveness

should show larger improvements in their ER-WR scores. In contrast, students in classes delivered in a less optimal way should demonstrate smaller ER-WR improvements.
The following sections examine the relationships between students’ average total ER-WR scores, and average responsiveness and quality ratings for each faculty member, at a macro level. Then assessment results for each element of the ER-WR rubric are integrated with specific features of the fidelity checklist and stratified by faculty participant, to examine this relationship at a more detailed level. Lastly, I profile the intervention implemented by the faculty participant whose students demonstrated the greatest improvements; this is perhaps the “best” version of the ER intervention. I also incorporate qualitative observations not captured in the fidelity checklist.

**Relationships among ER skills, responsiveness, and quality.** Figure 7 displays the relationships among average perceived student responsiveness, quality of intervention implementation, and students’ ER-WR average total scores at post-test stratified by faculty participant. As depicted in the graph, Faculty #4, #5, and #6 implemented the ER intervention with similar quality, overall; however, students in Faculty #4 and #5’s classes tended to earn higher ER-WR scores than students in Faculty #6’s class. The difference may be because students in Faculty #4 and #5’s classes tended to be more responsive.

On the other hand, Faculty #1 and #3 had higher perceived responsiveness, but students’ post-test ER-WR scores were lower, compared to students from Faculty #6 class. This is likely because Faculty #1 and #3 also received slightly lower quality of implementation ratings than Faculty #6. Thus, the positive influences of higher perceived responsiveness on students’ ER skills (i.e., ER-WR scores) may not alleviate the negative influences of lower implementation quality.
Figure 8 displays the same information as Figure 7, but in terms of faculty participant profiles. Both figures suggest an interaction effect of perceived student responsiveness and quality of implementation. For example, the effect of perceived student responsiveness on students’ upper-level ER skills (i.e., ER-WR scores) may be dependent on the quality with which the intervention was implemented. Or the possible interaction could be interpreted as: the effect of quality of implementation on students’ upper-level ER skills may be dependent on student responsiveness. Alternatively, as shown in Figure 8 by the blue line (i.e., Faculty #5), when both perceived student responsiveness and quality of implementation are high, there may be an additive, positive effect on students’ upper-level ER skills.

Thus, the effect of responsiveness and quality on students’ ER skills could be an interactive effect or an additive effect. It is not possible to discern this from the graphs. Given a larger sample size (e.g., more than $N = 6$ faculty participants), multiple regression analyses could be used to test whether the interaction effect was significant. At the least, both responsiveness and quality appear to be important contributors to students’ learning improvement. Moreover, the relationship between ER-WR skills, responsiveness, and quality appears to be positive, yet complex.

Figure 8 also visually depicts how Faculty #2’s students earned the second highest average ER-WR total scores, even though Faculty #2 had lower average perceived student responsiveness than other faculty participants. Faculty #2 had relatively high quality ratings though, which might have mitigated or compensated for the effects of lower student responsiveness. Alternatively, higher student responsiveness may not be able to compensate for implementing the intervention features with lower quality.
Consider Faculty #1 and #3; both tended to have higher perceived student responsiveness, on average, compared to Faculty #2. Yet, both Faculty #1 and #3 tended to implement the features with lower quality. Consequently, their students’ ER-WR average scores were lower compared to Faculty #2’s students’ scores (and the rest of the faculty participants as well).

As described previously, the students enrolled in the courses spanned the continuum from freshmen to seniors. The study also included a variety of different class types (e.g., lecture, seminar, required, non-required) (See Table 16). Therefore, perhaps the relationships shown in Figures 7 and 8 are bi-products or simply proxies of student developmental stage and/or class type. To examine this hypothesis, Figure 9 provides the same data shown in Figure 8, categorized according to class type. More specifically, the orange line represents a required, general education course that was comprised of mainly lower level students (i.e., first-year students) that primarily used lecture based pedagogies. The grey lines represent non-required courses (e.g., electives) that primarily used active learning or seminar based pedagogies, included a service learning component, and were comprised of mainly upper level students (e.g., third and fourth year students). Lastly, the purple lines represent courses that were required for a major, comprised of mainly upper level students, and primarily used lecture based pedagogies.

Figure 9 suggests that the relationships between the implementation fidelity data and assessment results were not merely due to students’ developmental levels (i.e., maturation) and/or class type. For instance, Faculty #2’s class (i.e., the orange line) was comprised of mainly lower-level students, but they earned higher ER-WR scores at post-test, on average, compared to other classes that were comprised of upper level students.
The courses taught by Faculty #3, #4, and #5 were categorized into the same class type (i.e., non-required class with service learning and mainly upper level students); however, their students did not earn similar post-test ER-WR scores. Indeed, Faculty #5’s students’ post-test ER-WR scores were nearly twice as large as Faculty #3’s students. Similarly, average quality ratings did not differ according to class type. Faculty #4 and #5’s average quality was more similar to Faculty #2 (i.e., orange line) and #6 (i.e., purple line), neither of which were classified into the same class type as they were (i.e., grey lines).

The categorizations shown in Figure 9 suggest that the fidelity data and assessment results were related for reasons beyond student developmental or maturation level and class type. Although there may be some effect of student developmental level and/or class type, this effect was not large enough to “overshadow” the effects of the ER intervention on students’ ER-WR scores. Recall the student demographic information discussed in Chapter Three. Student participants from the current study were slightly younger, on average, compared to students in the A-day spring 2016 comparison group. Yet, participant students demonstrated greater ER proficiency compared to the spring 2016 students.

Moreover, it is encouraging that the piloted ER intervention positively influenced students’ ER abilities regardless of developmental level or class type. In other words, the intervention appeared to have positive, aggregate influences on students of lower-level and upper-level maturation or development, in a more lecture-oriented class and in a more seminar or discussion based class, etc.

Additionally, the adherence component of implementation fidelity relates to students’ upper-level ER-WR skills. Notice that responsiveness, quality, and ER-WR
scores are all averages and are on similar metrics (i.e., ranging from 1 to 5 or 0 to 4, respectively). Adherence, however, is a frequency count (i.e., count of the number of times a specific feature was implemented) ranging from 31 to 90 times. To graphically compare quality, responsiveness, ER-WR scores, and adherence, all variables were converted to z-scores. Standardization put all of the variables on the same standard deviation metric (i.e., ranging from -1.50 to +1.75). Figure 10 displays the relationship between all of the fidelity variables and students’ average ER-WR post-test scores, in terms of faculty participant profiles.

Figure 10 has to be interpreted differently than Figures 7, 8, and 9 because the variables have been standardized; therefore, all comparisons within each variable are relative (i.e., or more specifically relative to the mean, so average ER-WR scores across all six faculty participants are compared relative to the grand mean ER-WR score, requiring some faculty to score above the mean and some below the mean). For example, Figure 10 shows that Faculty #2, #4, and #5’s students’ ER-WR scores tended to be above the mean improvement. Faculty #5’s students scored, on average, nearly 1.75 standard deviations above the mean, whereas Faculty #2 and #4’s students scored about half of a standard deviation or less above the mean, on average. Faculty #1, #3, and #6’s students’ tended to score below the mean. For responsiveness, Faculty #5, #4, and #3 tended to score above the mean, whereas the remaining faculty participants scored below the mean. Note that Faculty #2’s average student responsiveness ratings were about 1.50 standard deviations below the mean, on average. For quality, Faculty #3 and #1 tended to be below the mean by at least one standard deviation. The remaining faculty participants
were above the mean, with Faculty #4 at the highest (i.e., around one standard deviation above the mean).

Faculty #2, #3, and #5 tended to be above the mean for adherence. Although Faculty #5’s students earned the highest post-test ER-WR scores, Faculty #5 was only slightly above the mean in terms of adherence (i.e., about one fourth of one standard deviation above the mean). Meanwhile, Faculty #2 was nearly one and half standard deviations above the mean for adherence. Faculty #3’s adherence was nearly two standard deviations higher than Faculty #1’s, yet students from both of their classes demonstrated similar ER skills (i.e., scored similarly on the ER-WR). Recall, Faculty #1 and #3 had the lowest average quality ratings, relative to the mean quality rating across all the faculty participants. Thus, even though Faculty #3 was implementing the specific features noticeably more frequently than Faculty #1, the features may have been implemented with too low of quality to positively influence students’ ER abilities. Quality of implementation appears to be more salient to performance than quantity or frequency of implementation. Recall, Faculty #5 tended to implement the specific features fewer times relative to Faculty #3 and #2, but Faculty #5’s students earned higher ER-WR scores, comparatively. Perhaps because Faculty #5’s perceived student responsiveness and quality of implementation were higher than Faculty #3 and #2’s. Again, this suggests that higher student responsiveness and quality may have an additive, positive influence on students’ upper-level ER skills.

Although high enough quality of implementation may be able to compensate for lower student responsiveness, the reverse may not be true. Positive influences of high responsiveness may not mitigate the negative influences of low quality of
implementation. If the intervention is delivered with lower quality it might not matter how responsive or engaged students are. Although students may be actively “receiving” or “engaging” with the intervention, it is not being delivered to them in a high enough quality way to positively influence their ER abilities. This is suggested by the line representing Faculty #3 in Figure 8. Perceived student responsiveness was rated relatively high. However, quality of implementation was lower, relative to the other faculty participants. And thus, students’ average post-test ER-WR scores were also lower compared to other faculty.

Furthermore, quality and responsiveness together may have greater influence on students’ ER skills than frequency of adherence. Relatively speaking, Faculty #5 tended to implement the features with just over average frequency, but his/her students’ ER-WR scores were over one standard deviation higher than any other faculty participants’ students, on average. Faculty #5 had above average responsiveness and quality, relative to the other faculty participants, which may have additively or jointly influenced their students’ ER skills.

Overall, the relationships between ER-WR skills, responsiveness, and quality appear to be positive, although complex. With quality potentially offsetting some of the negative influences of low responsiveness on students’ ER abilities. Frequency of adherence tended to have less positive effects on students’ ER abilities compared to responsiveness and quality. To understand the complex relationships between implementation fidelity data and assessment results requires a more in-depth examination of each rubric element and intervention specific features.
**Element A: identify ethical issue in its context.** Element A of the ER-WR rubric was most aligned with the following specific features of the fidelity checklist (See Appendix F):

- Read/review rubric
- Provide/discuss example of a decision making process w/AND w/out ethical reasoning
- Identify and explain how characteristics or features make the case (in)effective
- Review/build a “strong” or “effective” example of ethical reasoning

As shown in Table 18, none of the faculty participants implemented the “Identify and explain how characteristics or features make the case (in)effective” specific feature whatsoever. But Faculty #5 and #6 implemented the other three specific features the most compared to other faculty participants. Interestingly, the students who completed Faculty #4’s course demonstrated the most growth from pre- to post-test in Element A (See Table 11). Also, Faculty #4 did not implement the “Review/build a “strong” or “effective” example of ethical reasoning” feature at all; therefore, it may not be salient for increasing students’ abilities to identify an ethical issue in context.

Across all students, post-test ER-WR scores tended to be the highest for Element A (See Table 11). Recall, previous chapters discussed that Element A might just be easier than the other elements. At post-test, the students who completed the course taught by Faculty #5 scored the highest on Element A, comparatively. Student responsiveness was higher for the implementation of these two features for Faculty #4 and #5 compared to Faculty #6. Thus, although Faculty #6 was implementing the specific features just as frequently as the other faculty members and with decent quality, lower average student
responsiveness may indicate that students were not actively “receiving” or engaging enough to make a difference in their ER skills assessed by Element A. Compared to Faculty #6, Faculty #4 and #5 taught smaller class sections perhaps making easier small group work (i.e., with ~3-4 students), peer-to-peer interaction, and engagement. Faculty #1 did not implement any of these features; his/her students tended to demonstrate the lowest gains in Element A from pre to post-test.

The fidelity and assessment data seem to suggest that the first two features (Read/review rubric and Provide/discuss example of a decision making process w/AND w/out ethical reasoning) may be salient to the ER intervention (i.e., given Faculty #1’s students showed no gains, but other Faculty participants’ students did). These two features should be implemented at least a couple of times as part of the intervention, but the implementation must be done such that students are highly responsive. There appears to be a relationship between implementation fidelity and students’ upper-level ER abilities (i.e., ER-WR scores). Students who experienced none of the specific features aligned with Element A tended to show the smallest improvements in their Element A scores, over time. Student responsiveness and quality may moderate the relationship between implementation fidelity and students’ essay scores. When specific features are implemented very frequently – but at lower levels of student responsiveness and/or quality – there appears to be a less meaningful relationship between implementation fidelity and improvements in students’ upper-level ER abilities.

**Element B: 8KQ reference.** Element B of the ER-WR rubric was most aligned with the following specific features of the fidelity checklist:

- Read/review rubric
• Elaborate or unpack the 8KQ
• Students experience a “check point”
• Map 8KQ to some other work
• Critique/edit/comment/annotate the 8KQ
• Review/refresh 8KQ
• Process something (debate, case, discussion, etc.) using 8KQ

The students who completed the course taught by Faculty #5 demonstrated the most growth in their ER-WR scores for Element B and had the highest post-test scores for Element B (See Table 11). However, Faculty #5 did not implement these features more frequently than all of the other faculty. For example, Faculty #2, #3, and #6 implemented these features more frequently (See Table 18). But Faculty #5 had higher perceived student responsiveness across these features, compared to the other faculty members. During class, Faculty #5 asked students to participate in activities that required them to physically move around the room, stand up, throw and catch a ball as a self “check” of their understanding of the 8KQ, work in small groups, etc. Again, perhaps student responsiveness is just as important (if not more important) than the frequency with which the features are implemented. Faculty #5 also implemented the features with just as high or higher quality compared to the other faculty participants.

The students who completed the course taught by Faculty #1 tended to show the least gains, over time, for Element B. Faculty #1 implemented these features less frequently than the majority of the other Faculty participants, and their students demonstrated the least improvement from pre- to post-test. Faculty #1 tended to implement the ER intervention towards the end of class; sometimes running out of time.
to fully implement the intervention as she/he may have intended. Also, given the larger class size and physical space limitations of the classroom, it was difficult for students to form smaller working groups to analyze case studies, engage in reflection, etc. For example, students often had to form larger groups of ~6-7 students, packed in tightly with other groups; some students did not have to (or did not get a chance to) contribute to the group conversations/discussions. Note that the “Critique/edit/comment/ annotate the 8KQ” was only implemented once (by only one faculty participant: Faculty #1); perhaps this feature is not important for increasing students’ skills related to Element B or faculty require more training in order to implement this feature.

In sum, the integrated results suggest that all but one of the specific features aligned with Element B may be salient to an ER intervention that promotes students’ abilities to reference the 8KQ (i.e., Element B). But student responsiveness and quality of implementation may be more paramount than frequency of implementation. Moreover, fidelity of implementation appears to be (at least somewhat) related to improvements in students’ ER-WR scores. Students who experienced the specific features, with higher quality of implementation and higher responsiveness, may also be the same students who earned higher ER-WR Element B scores at post-test.

**Element C: 8KQ applicability.** The following specific features of the checklist most aligned with Element C of the ER-WR rubric:

- Elaborate or unpack the 8KQ
- Read/review rubric
- Identify where/how each of the 8KQ are/ are not applied within the case
- Give/discuss rationale for how each of the 8KQ are/are not applied
Across all faculty participants, the “Identify where/how each of the 8KQ are/are not applied within the case” and “Give/discuss rationale for how each of the 8KQ are/are not applied” specific features were implemented more frequently than the other two features.

Students who completed the course taught by Faculty #5 demonstrated the most growth in their ER-WR scores for Element C and had the highest post-test scores for this element (See Table 11). Faculty #5 implemented all four of these specific features at least once. She/he also implemented these four features more frequently than four of the other faculty participants. Faculty #2 implemented these specific features the most frequently; their students demonstrated the second largest improvements in Element C scores. Faculty #2 provided in-depth elaboration for all 8 of the KQ. She/he also integrated other philosophies and theories to help bolster students’ deeper understanding of the 8KQ. During class, Faculty #2 asked students to grapple with several different kinds of ethical situations, from medicine (e.g., use of Growth Attenuation Therapy to stunt growth of disabled child) to larger philosophical or societal contexts (e.g., English criminal case involving survival cannibalism). Faculty #2 mainly implemented these features in a more lecture-based way. Students typically remained in their seats, seldom asked questions, rarely discussed or processed material with other students, etc. The main difference between Faculty #5 and #2’s classes was responsiveness. Perceived student responsiveness was rated higher for Faculty #5 (i.e., 5 on average) for the implementation of these features than for Faculty #2 (i.e., 3 to 4 on average). Even though Faculty #2 implemented these four features more frequently and with very slightly higher quality, on average, the students from Faculty #5’s class were rated as being more responsive while
these specific features were implemented. Perhaps responsiveness made a difference for enhancing students’ abilities to apply the 8KQ (i.e., Element C).

Faculty #1 and #6 implemented these four features the least frequently compared to the other faculty participants. The students who completed Faculty #1’s class demonstrated the least improvements in their Element C scores. Given Faculty #2 and #5 implemented these features the most frequently, with medium to high quality, and their students tended to demonstrate greater improvements in Element C scores. Perhaps all four of these features are salient to an effective ER intervention. However, it may be important to implement the “Identify where/how each of the 8KQ are/are not applied within the case” and “Give/discuss rationale for how each of the 8KQ are/are not applied” features more frequently than the other two features. Similar to patterns observed for Elements A and B, ensuring that students are engaged during implementation may be just as crucial as frequency of implementation. Fidelity data suggest a relationship between implementation of the ER intervention and students’ ER-WR scores for Element C.

**Element D: analyzing individual KQ.** Element D was most aligned with the following fidelity checklist specific features:

- Read/review rubric
- Experience (visually or another sense) the 8KQ analysis processes
- Experience some analysis (or breaking a part) of at least 1 KQ
- Identify obstacles or pitfalls to analysis
- Consider contextual factors
• Expose/demonstrate/suggest how multiple perspectives compete/interact within same KQ

• Multiple stakeholders and/or perspectives are identified or considered

Faculty #3 implemented these seven features most frequently compared to the other faculty members. However, their students did not demonstrate the largest improvements; Faculty #5’s students did. The students who completed the class taught by Faculty #5 also earned the highest scores on Element D at post-test. On average, quality of implementation and perceived student responsiveness were higher for these seven features for Faculty #5 compared to Faculty #3. Similarly, Faculty #2 implemented these features with slightly greater frequency than Faculty #5, yet their students did not demonstrate as much improvement. Quality and perceived student responsiveness may explain why.

For example, student responsiveness was higher, on average, for Faculty #5 compared to Faculty #2. Faculty #5 also tended to implement these features with just slightly higher quality, on average, compared to Faculty #2. For instance, in Faculty #5’s class, students were consistently prompted to consider multiple stakeholders. Given the disciplinary area of Faculty #5’s class, it seemed easy and intuitive for students to identify and consider multiple perspectives within a particular ethical situation. Because Faculty #5 used case studies and examples that were very applicable to students’ current placements (and to their future careers) students were able to identify and consider contextual factors as they applied the 8KQ. Students did not have to imagine or conjecture what the contextual factors might be for a given ethical situation or what perspectives might need to be considered. Rather, because the case studies were so
applicable to their current experiences and disciplinary contexts, students could discuss and reflect on real-life, tangible perspectives and contextual parameters. Implementing the Element D features using “close to home” (e.g., highly applicable, palpable) case studies, examples, and situations may have helped Faculty #5’s students practice and improve their ER skills.

Faculty #1 implemented these features the least frequently (i.e., only implemented one of these specific features in their class). Their students demonstrated the least improvement, on average, for Element D. Faculty #4 and #6 implemented these features about half as frequently as Faculty #2, #3, and #5. Yet, Faculty #4’s students, on average, demonstrated almost as much improvement as Faculty #5’s students. Given Faculty #4 and #5 implemented these features with similar average quality and perceived student responsiveness, perhaps quality and responsiveness are more salient than frequency of adherence. In other words, Faculty #5 implemented the Element D features with greater frequency than Faculty #4, but did not see a much larger magnitude of positive influence on her/his students’ ER-WR scores. Quality and responsiveness were similar for Faculty #4 and #5; thus, their students’ ER-WR scores improved similar amounts.

Recall, Faculty #4 taught a seminar based class with a smaller number of students. During class, students worked in small groups to analyze ethical situations, shared their discussions and analyses with the class, and even debated/critiqued other groups’ analyses. All of these activities required fairly high levels of student engagement or responsiveness. Faculty #4 also had students actually write out their 8KQ analysis, under a specific time constraint, during class. Then students used the ER-WR rubric to evaluate other students’ 8KQ analyses. These activities engaged students while also allowing them
to practice ER skills in the same way they would be assessed (i.e., writing an essay in response to the ER-WR prompt), and enhancing their familiarity with the ER-WR rubric. The pattern of findings aligns with results from previous elements: student responsiveness and quality of implementation matters, just as much if not more than frequency.

**Element E: weighing relevant factors and deciding.** The following specific features most aligned with Element E:

- Read/review rubric
- Identify/discuss which (if any) aspects of the case are “compelling?”
- Arrive at or grapple with particular conclusion or decision point

Students who completed the class taught by Faculty #5 demonstrated the greatest pre- to post-test improvement in their Element E scores. At post-test, Faculty #5’s students also earned the highest scores on Element E.

Other than Faculty #1, the faculty participants tended to implement these three features with similar frequency. Yet, improvements in their students’ abilities to weigh relevant factors and make a decision varied. Improvements ranged from a 0.143 pre- to post-test increase for Faculty #3, to a 1.149 increase for Faculty #5. Again, perceived student responsiveness was slightly higher, on average, for Faculty #5 compared to the other faculty members. But quality of implementation was lower for Faculty #5.

Faculty #2’s students had slightly lower responsiveness, a pattern observed for other elements. In addition, Faculty #3 and #5 had slightly lower quality, and Faculty #1 implemented fewer features, less frequently. Faculty #3 implemented the features aligned with Element E more frequently than Faculty #1, yet their students tended to demonstrate
less improvement in their abilities to weigh the 8KQ and decide. Perhaps this was due to the fact that Faculty #1 tended to implement these features with slightly higher quality than Faculty #3. Perceived student responsiveness was highest for Faculty #5’s students; supporting previous findings that student responsiveness may be just as imperative as frequency of implementation. While findings for Element E suggest a relationship between naturally occurring differences in fidelity of implementation and assessment data, the relationship may not be as apparent as it was for Elements A through D.

Examination of assessment results for each rubric element, integrated with features of the fidelity checklist, suggested a relationship between fidelity of implementation and students’ upper-level ER abilities. Observationally, the relationship appears to be positive. However, the relationship is not as simple as: the more frequently faculty participants implemented specific features, the more positive influence they had on students’ ER skills.

Rather, the relationship between implementation fidelity and assessment performance may be moderated by student responsiveness and by quality of implementation. When perceived student responsiveness and/or quality are higher, the relationship between implementation fidelity and students’ upper-level ER skills appears to be stronger. Student responsiveness and quality may have an additive positive effect on students’ ER skills. It is advantageous to further examine the intervention delivered to the students who improved their ER-WR skills the most. The following section profiles the intervention for Faculty #5’s class.

**Intervention profile for class that demonstrated the most improvement over time.** Students who completed the course taught by Faculty #5 tended to show the
greatest improvements in their ER skills. In addition, these students tended to earn higher total ER-WR scores at post-test compared to students in other classes. Faculty #5’s class was noticeably different from the other classes in terms of learning improvements, posttest ER-WR scores, student responsiveness, and (to a lesser extent) quality.

To make Faculty #5’s class more tangible, I profile the intervention that these students received. I highlight specific features and characteristics that distinguish Faculty #5’s intervention from the interventions implemented in the other classes. The profile can inform successive modifications to the fidelity checklist and help other faculty implement well-aligned ER interventions in their own classes.

Holistically, Faculty #5 implemented the checklist features equally. Other faculty participants implemented certain features with very high frequencies, while implementing others with much lower (or no) frequency (e.g., Faculty #2). Note, Faculty #6 also implemented the checklist features fairly equally, but their students’ responsiveness was not as high as Faculty #5. Also, Faculty #5 did not implement any particular program component more frequently than all of the other faculty participants. Therefore, it does not appear that Faculty #5 emphasized or prioritized any one particular subsection of the fidelity checklist. Results suggest that at least one of the specific features from each program component is salient to improving students’ ER skills.

As discussed previously, the high levels of student responsiveness during intervention implementation differentiated Faculty #5’s class from the others. Recall, there were instances in which Faculty #5 implemented certain features less frequently than other faculty, yet her/his students still earned higher scores on the ER-WR rubric elements. It appears that Faculty #5 prioritized student engagement over frequency of
implementation. She/he tended to implement certain features less frequently, but ensured that when she/he did implement a feature it was done so in an engaging way. Faculty #5 consistently contextualized the 8KQ and ER activities, demonstrations, etc. within the course content or area of study. She/he connected case studies and other 8KQ activities to specific, tangible instances that her/his students were currently experiencing in practicums/placements. Faculty #5 also connected case studies to instances that students would experience in their future jobs. Overall, Faculty #5 consistently provided contexts that conveyed the usefulness and importance of ER skills to the specific field that his/her students were studying. Real-world, contextualized examples likely enhanced student engagement.

Faculty #5 used a variety of tools or methods to implement the various checklist features, which further differentiated this class from others. Although Faculty #5 did incorporate some of the same tools as other faculty participants (e.g., use of case studies), she/he did not rely on one method to deliver the specific features. Faculty #5 used the greatest variety of activities, exercises, and so forth. To help students review/internalize the 8KQ and what each represented, Faculty #5 brought a bouncy ball to class and had students throw it around the room to one another. The student who caught it had to state one of the 8KQ from memory and define it. If she could not remember the definition, then she had to pass the ball to another student who would provide the definition. To help students experience and practice the 8KQ analysis process, during one class activity, Faculty #5 had students select one item from a large pile of random items and then create a metaphor using that item to represent the 8KQ ER process. For another activity, Faculty #5 had students analyze an ethical dilemma and physically “weigh and balance” the 8KQ.
Students used wooden blocks to represent the various 8KQ. Larger or heavier blocks represented KQ that were more applicable to the dilemma, while smaller blocks represented KQ that were less applicable. KQ that were not applicable were not represented by any blocks.

Faculty #5 also asked students to analyze discipline-based case studies using the 8KQ. These case studies were highly applicable to the students’ field of study. Case studies involved ethical scenarios and situations that the students would have to grapple with in their current placements and future jobs. Other faculty participants also used case studies as a pedagogical tool to implement the intervention specific features. Yet, those case studies were not always quite as overtly applicable to students’ current or future jobs as the case studies used by Faculty #5. Whenever students discussed the 8KQ, Faculty #5 prompted them to consider and “unpack” multiple perspectives. If student groups were using the 8KQ during an activity in class but they were excluding a particular KQ (e.g., Rights), Faculty #5 would spend additional time on the excluded KQ, explaining how it could be applicable.

Faculty #5 did not implement the following specific features:

- Read/ Review SLOs
- Critique/edit/comment/annotate the 8KQ
- Identify/discuss which (if any) aspects of the case are “compelling”
- Identify and explain how characteristics or features make the case (in)effective, and
- Identify obstacles or pitfalls to analysis
Without experiencing these features, Faculty #5’s students still improved their upper-level ER skills. Perhaps these specific features are not salient to improving students’ ER abilities? Alternatively, Faculty #5’s students might have demonstrated even greater improvements if they had experienced some, or all, of the aforementioned features. For instance, Faculty #5’s students demonstrated the smallest improvements in their Element E rubric scores, compared to their scores for Elements A through D. The “Identify/discuss which (if any) aspects of the case are ‘compelling’” specific feature was aligned with Element E.

Given I attended nearly all of the classes to collect fidelity data, I have qualitative observations about each class not captured within the specific features of the fidelity checklist. It is worth noting that Faculty #5’s class was characterized by a very collegial group of students. While observing this class, I noted a great sense of trust and respect among the students. The class culture was overwhelmingly one of active learning, which may help explain why the students in this class tended to be highly engaged with the intervention features. And this engagement may have positively contributed to improving these students’ ER skills throughout the semester. Furthermore, Faculty #5 demonstrated understanding and mastery of a variety of pedagogical techniques. Techniques that she/he used to implement the specific features of the intervention. I think that Faculty #5’s pedagogical efficacy was a unique factor that perhaps made the intervention implementation more successful, comparatively.

**Summary.** Overall, naturally occurring differences in implementation of the ER intervention appear to be related to student performance on the ER-WR. Indeed, student responsiveness and quality may moderate the relationship between intervention
implementation and performance on the ER-WR. Students who experienced the majority of the specific features with high responsiveness and quality, tended to demonstrate greater improvements in their higher-level ER abilities.

The intervention delivered by Faculty #5 tended to be the most effective, comparatively. However, remember that every faculty participant implemented an intervention that positively influenced their students’ ER skills, to varying degrees. That is, all faculty participants implemented the piloted intervention with moderate to high fidelity. And their students tended to demonstrate improvements in their ER abilities, on average. Moreover, the students who completed a class taught by Faculty #2, 4, or 5 earned ER-WR scores above JMU’s strategic plan goal of a 2.0, on average (See Table 11).

Results from integrating assessment and fidelity data have implications for future implementations of the ER intervention. It may be important to use a variety of pedagogical techniques or teaching tools when implementing the checklist features (like Faculty #5 did). Additionally, certain checklist features may not be salient to improving students’ ER skills:

- Read/ Review SLOs
- Critique/edit/comment/annotate the 8KQ
- Identify/discuss which (if any) aspects of the case are “compelling”
- Identify and explain how characteristics or features make the case (in)effective,
- Identify obstacles or pitfalls to analysis.

Alternatively, faculty may need further development or training to successfully implement these features. Faculty may need help to create class activities, assignments,
presentations, demonstrations, etc. that align with each of the aforementioned features.

Faculty participants struggled to understand what it meant for a case to be “compelling” and how they could subsequently teach this to their students. The following chapter provides more detailed recommendations for intervention modifications, study limitations, and implications for higher education institutions beyond JMU.
CHAPTER 5

Discussion

Just as Apple integrated MP3 players, music management software, and broadband internet capabilities in the early 2000’s, I integrated assessment methodology with curriculum and pedagogy, under the framework of a learning improvement model (Fulcher et al., 2014). Specifically, researchers used implementation fidelity methodologies to create and empirically study an ER intervention (i.e., curricula and pedagogies). As described in Chapter Four, integrating assessment components with teaching/learning components helped faculty demonstrably improve students’ ER skills. Improvements were evidenced by statistical significance, a large standardized effect (i.e., $d = 1.7$), and a meaningful practical effect (i.e., students moved from “Developing” to “Good” on the ER-WR rubric). On a broader note, the study successfully bridged the disconnect between learning outcomes assessment and student learning improvement for an important skill area: ethical reasoning (AAC&U, 2013).

The following sections describe study implications specific to the institution where the study took place (i.e., James Madison University), followed by broader implications for higher education. For stakeholders at JMU, I provide suggestions for modifications to the ER intervention checklist and strategies for scaling the intervention up to the program and university levels. Specific study limitations and future directions for research are also considered.

Specific Implications for JMU

JMU demonstrated learning improvement through an ethical reasoning project. Keep in mind, such examples of learning improvement are rare in higher education...
(Banta, Jones, & Black, 2009; Banta & Blaich, 2011). This study was the first of its kind conducted at JMU. That is, previous studies did not apply the Simple Model of improvement, implementation fidelity methods, and outcomes assessment to a semester-long intervention for an academic program. Through this work, JMU faculty and stakeholders addressed three major university-level needs:

- Identify and describe an ER intervention that will be effective enough to help students achieve the University strategic plan standard of a 2.0 on the ER-WR rubric by the year 2020;
- Empirically study the ER intervention to determine the fidelity with which it was implemented, the extent to which it positively influenced students’ ER abilities, and which features of the intervention may be salient for ER learning improvement; and lastly,
- Ensure the intervention is flexible (e.g., adaptable) enough to be implemented in general education and major-specific classes, and scaled up across the university.

Faculty participants were able to create a flexible ER intervention, and then implement various features of that ER intervention (See Appendix F) with moderate to high fidelity, across different disciplines and class types (See Table 16). Features of the intervention positively influenced students’ ER skills, across disciplinary contexts and various student developmental levels (See Figure 9). Students in three of the six classes were able to exceed the university strategic plan goal of a 2.0 on the ER-WR rubric. Thus, results suggest that faculty participants created an ER intervention that could help students, on average, achieve (or even surpass) the university-level goal. Now, JMU stakeholders can
disseminate the specific features of that intervention to help promote effective, evidence-based ER education, at the program and university-levels.

Improvements in students’ ER skills were evidenced at the individual course levels; however, the Simple Model emphasizes program-level learning improvement. Therefore, Madison Collaborative leadership and faculty members can use study results to scale-up the ER intervention to the program and university-levels. When engaging in this important task, stakeholders should first consider the following recommendations.

**Recommendations for modifications to the ER intervention checklist.** As discussed in Chapter Four, the following intervention specific features need to be investigated further:

- Read/ Review SLOs
- Critique/edit/comment/annotate the 8KQ
- Identify and explain how characteristics or features make the case (in)effective
- Identify/discuss which (if any) aspects of the case are “compelling,” and
- Identify obstacles or pitfalls to analysis

Across all faculty participants, the first three features listed above were implemented rarely or never (See Table 18). Faculty #5, for example, did not implement any of these five features. Perhaps the five features are not salient to the intervention, given they were implemented rarely or infrequently, but students’ upper-level ER skills still demonstrated improvements over time. Alternatively, if faculty participants would have implemented the five features more frequently, then improvements in students’ ER skill may have been even greater in magnitude. Fortunately, MC leadership can further investigate this finding through future research.
If MC leadership and faculty decide to retain the five features as part of the intervention checklist, they would need to determine the extent to which the features have positive effects on students’ learning (i.e., through future research studies). That is, faculty would need to actually implement the features to assess their efficacy. MC stakeholders may need to provide additional training to help faculty implement the features. Note, during the creation of the intervention checklist, faculty struggled to understand characteristics that make a case “compelling.” Imagine, in future studies, that faculty were able to implement the five features with high fidelity. Yet, students’ ER skills did not demonstrate improvements of larger magnitude than participant students from this study. Results would suggest that the five features may not be salient to improving students’ ER skills; the features could be removed from the intervention checklist.

Additionally, fidelity researchers noticed overlap among some of the intervention specific features when they were applying the fidelity checklist to collect data. For instance, the “Students process something (debate, case, discussion, etc.) using 8KQ” specific feature tended to overlap or be redundant with several of the features in the “Case Study” program component including: “Identify where/how each of the 8KQ are/are not applied within the case,” “Give/discuss rationale for how each of the 8KQ are/are not applied” and “Engage in reflection (e.g., could be formal or informal, written, oral, group, what issues did you have, what was easy/hard).” Sometimes this overlap made the checklist difficult to use (e.g., it was difficult to record duration times for specific features without “double counting” time due to the overlap). Other features such as “Expose/demonstrate/suggest how multiple perspectives can compete/interact w/one
another within the same KQ” and “Multiple stakeholders and/or multiple perspectives are identified or considered” also appeared to be redundant, and may require further differentiation or clarification.

To further improve the checklist (and thus the ER intervention), MC leadership could work with faculty members to differentiate these overlapping features. Or if the features cannot be distinguished from one another, then perhaps they are redundant and some of them could be removed from the checklist (e.g., excluded from the intervention). Disentangling the overlapping features should help fidelity researchers use the checklist more effectively and consistently. On the other hand, having redundancies in the checklist could be pedagogically useful. The overlap could help faculty reinforce concepts or application in a more intentional way. If the MC leadership decides that redundancies are pedagogically useful, they should create strategies for collecting accurate and consistent fidelity data despite overlap across the features.

Results suggested that quality and student responsiveness are important characteristics of intervention implementation. More specifically, high quality of implementation coupled with high student responsiveness led to the greatest improvement in ER skills. Furthermore, high implementation quality positively influenced students’ upper-level ER skills, so much so that quality may be able to compensate for lower student responsiveness. Faculty who want to implement the ER intervention in their classes should be aware of these findings. Ultimately, if faculty cannot implement all specific features, those they do implement should be done so with high quality and in ways that promote high student responsiveness. One faculty participant commented that responsiveness seemed to be higher in his course during
spring 2017 compared to fall 2016 (i.e., when fidelity data were collected). Certainly, responsiveness will vary somewhat from one semester to the next, depending on the students enrolled in the class. In addition to the affect and attitudes that students bring into the classroom, responsiveness could also be an attribute of the faculty member. And thus, faculty can promote responsiveness through their own demeanor and teaching strategies. Through the current study, faculty participants have provided examples of pedagogical techniques that can promote high student responsiveness.

Faculty participants, like Faculty #5, were able to implement several of the specific features in high quality ways, while also promoting high perceived student responsiveness. Perhaps MC leadership could ask Faculty #5 to share specific resources or strategies (e.g., pedagogical tools/mechanisms, class activities, etc.) with other faculty members who are interested in teaching ER. Some of the strategies used by Faculty #5 were simple, and could (theoretically) be easily adapted into various classes. Recall, Faculty #5 used activities that required students to move around the room, use physical objects to promote their understanding and application of the 8KQ, and so forth.

Other faculty participants also used noteworthy strategies; most of which would be easily accessible for other faculty to apply in their classes – regardless of the disciplinary context or class type. Faculty #4 and #6 asked students to create their own case study examples and then explain/act them out. Classmates then had to differentiate “difficult decisions” from “ethical situations.” Faculty #6 also asked students to role play different ethical situations and describe various 8KQ by creating a visual aid. These represent types of activities and demonstrations that promote student responsiveness. And faculty may be able to adapt these for use in their class, with minimal effort or difficulty.
Recommendations for continued faculty development and support. I also encourage MC leadership to continue providing faculty development opportunities. As mentioned previously, faculty need additional training to implement certain intervention features (e.g., “Identify/discuss which (if any) aspects of the case are ‘compelling’”). Moreover, faculty may need training to understand appropriate and effective uses of the intervention fidelity checklist.

I recommend that MC leadership partner with program/department heads and/or the campus faculty development center to provide faculty training opportunities and support. Partners from the campus assessment center could help educate faculty about implementation fidelity research and the intervention fidelity checklist. The MC could potentially offer mini-grants to help faculty implement fidelity research studies and longitudinal data collection methods (i.e., assess, intervene, re-assess). Also, programs could leverage pre-existing resources and infrastructures at the university to collect longitudinal data (e.g., University-wide Assessment day, etc.). Importantly, faculty participants from the current study would be exceptional training facilitators, as well as resources, for other faculty.

ER would be more pervasive at JMU if more faculty were to implement the effective strategies detailed in the intervention checklist, across a range of major and general education classes. Scaling up, however, is challenging. For instance, it requires faculty to collaborate with colleagues as they implement intervention features across multiple classes, within the same major or general education area. In the following section, I provide strategies to support scaling up.
Strategies for scaling up the ER intervention at JMU. Using the Simple Model framework (Fulcher et al., 2014), results demonstrated student learning improvements at the course level. Improvements were linked to specific curricula and pedagogies via the MC fidelity checklist (See Appendix F). Now, MC stakeholders and faculty may need to consider strategies for scaling this successful learning improvement initiative up to higher organizational levels (e.g., more tightly aligning it with the best practices put forth by the Simple Model).

It is important to ensure that the ER intervention is “scalable” because the goal of the MC is to positively influence all undergraduate students’ ER abilities, not just the students who are fortunate enough to have enrolled in one particular class that teaches the 8KQ. Furthermore, research suggests that ER skills require effortful development (Kohlberg, 1977), which should likely occur over the course of multiple classes. Study results demonstrated that one semester-long course could have positive effects on students’ ER abilities; however, there is still room for further improvements. Across all six classes, students scored 2.0 out of 4.0 on the ER-WR rubric, on average. Imagine if students experienced the piloted intervention (implemented with high fidelity) across multiple classes and contexts, rather than in just one course for one semester. Perhaps faculty could move students from being “Good” to “Extraordinary” ethical reasoners (i.e., from earning a score of 2.0 to 4.0).

To this end, the ER intervention needs to be implemented across multiple courses within a program (i.e., scaled up to the program level). Then the intervention needs to implemented across multiple programs and General Education classes (i.e., scaled up to
the university-level). One trajectory for scaling up to the program- and university-level would involve the following steps:

1. Focus on several programs who are willing to pilot the ER intervention within a class or two
   a. Collect fidelity data to determine the fidelity with which the intervention was implemented in those classes
   b. Collect outcomes assessment data to demonstrate ER learning improvements
   c. Use results to refine/improve future iterations of intervention implementation

2. Given Step 1 is successful, help the program implement the refined ER intervention across multiple courses within the program (i.e., scale up to the program level)
   a. Collect fidelity data to determine the fidelity with which the intervention was implemented across the program
   b. Collect outcomes assessment data to demonstrate ER learning improvements at the program-level

3. Given success at Step 2, pursue additional programs using similar strategies. As a greater number of programs implement the ER intervention across numerous classes within the program, the intervention will begin scaling up to the university level

4. Given success at Step 3, focus on one or two General Education Areas (e.g., Clusters) who are willing to pilot the ER intervention within a class or two
a. Collect fidelity data to determine the fidelity with which the intervention was implemented in those classes

b. Collect outcomes assessment data to demonstrate ER learning improvements

c. Use results to refine/improve future iterations of intervention implementation

5. Given success at Step 4, pursue any remaining General Education Areas using similar strategies. Once all General Education Areas implement the ER intervention, it will be fully scaled up to the university level.

The MC leadership now has initial validity evidence for the piloted ER intervention, which could enhance intervention “scale-ability.” Through design, the intervention was expected to have a larger effect on students’ upper-level skills. Results via the ER-W were consistent with this hypothesis, suggesting that the intervention positively influenced the particular type of ER skills it was intended to improve. Such evidence may convince otherwise skeptical faculty that partnering with the MC and implementing the ER intervention is worth their time.

I also suggest that MC stakeholders partner with the campus assessment and faculty development centers to provide additional faculty development opportunities. Implementation fidelity research and use of the fidelity checklist take time to learn. If faculty understand what the specific features mean and how they can positively influence students’ ER skills, they may be more likely to implement the intervention. Furthermore, the more comfortable faculty become with the MC fidelity checklist, the more faithful they will be to the effective ER intervention.
Participating faculty should also be encouraged to fill out the fidelity checklists for themselves. These data can serve as “self-audits”, prompting them to implement with high fidelity. Checklists completed by faculty members could also provide more exact “dosage” data for the Madison Collaborative. Currently, the MC collects relatively crude “dosage” data from students during University-wide assessment days every spring semester. Students are asked to recall and self-report the level of “exposure” to the 8KQ, quantified in hours, that they have previously received in general education and major classes. Although such data are useful, checklist fidelity data would provide more accurate information.

Critics may suggest that faculty will resist such “pre-packaged” or “pre-determined” interventions, which could threaten faculty’s sense of academic freedom. To address this concern, the MC should ask faculty participants to provide example assignments, demonstrations, lectures, activities, etc. that demonstrate how they integrated the intervention features into their courses. Faculty-sourced examples would showcase the flexibility or adaptability of the intervention features. Faculty of varied backgrounds, teaching different kinds of classes, across various disciplines, were able to implement intervention features successfully.

Note, faculty participants should be considered important liaisons between the MC and other faculty. Their “insider” knowledge of the intervention could motivate other faculty members to join the initiative. Faculty participant involvement will be an essential component of scaling up the intervention.

Scaling up also requires support from administrators who can provide organizational resources. For instance, deans or department heads could provide stipends
for multiple faculty members to attend a weeklong course-redesign process. There, faculty could integrate the ethical reasoning intervention features into their own courses. Working collaboratively, faculty could crowd-source ideas for assignments, activities, demonstrations, etc., or use resources produced through this study. Faculty could also discuss scaffolding of assignments, activities, demonstrations, etc. across various classes within their major or department. Given the flexibility of the intervention features, faculty could create multiple options for implementation, yet still work collaboratively across courses within the same program/department. The course redesign approach may be the most efficient way to scale up the intervention to the program and university levels, leveraging infrastructure and/or resources already in place at JMU.

**Study limitations.** JMU stakeholders have demonstrated that students can become “Good” ethical reasoners. Indeed, for the first time ever, students reached the university strategic plan goal for ER skills (i.e., average score of 2.0). In addition, implementation fidelity research provided a detailed ER intervention and empirical examination of the intervention’s efficacy. Results suggested that the intervention positively influenced students’ ER abilities. Stakeholders can demonstrate that learning improved *and* explain *why*. That is, the curricula and pedagogies detailed in the intervention specific features contributed to learning improvements. Nevertheless, there are important limitations that should be addressed.

First, consider the balance between internal and external validity. Internal validity can be defined as the extent to which inferences reflect a causal relationship between variables in the study (Shadish, Cook, & Campbell, 2002). Given this study was quasi-experimental (e.g., students were not randomly assigned to classes or majors or
professors, etc.), internal validity was imperfect. Although the study’s longitudinal design demonstrated that an ER intervention (i.e., the IV or manipulated variable) preceded improvements in students’ ER skills (i.e., the DV) and that implementation of the ER intervention was related to improvements in students’ ER skills, it did not (and cannot) rule out all other plausible explanations for this relationship. For example, there were several explicit threats to internal validity that could (at least partially) explain the relationship between the ER intervention and improvements in students’ ER skills, including:

- **Attrition**: some students did not complete the assessments at both pre- and post-test, some students enrolled in the faculty participants’ courses did not complete the assessments at all;
- **Testing**: completing the assessment instruments at pre-test could have affected students’ scores when they completed the assessments again several weeks later at post-test;
- **Selection**: because students were not randomly assigned to classes or randomly assigned to receive or not receive the ER intervention, it is possible that the average student who received the ER intervention differed in some important way from students who did not receive the intervention; and
- **Maturation**: students may have naturally improved their ER skills absent the ER intervention, perhaps because they grew older or smarter as a result of simply being at college for a greater amount of time (Shadish, Cook, & Campbell, 2002).

Without further study, it is difficult to ascertain how attrition or testing effects have influenced the results; both are potential limitations of the current study.
Selection threats to internal validity were alleviated by the study’s longitudinal design. For example, even if the participant students differed from other non-participant students in some relevant way, results still suggested improvements in ER skills. That is, at the least, the intervention positively affected one type or group of students. Selection threats would have substantial impact on internal validity for the cross-sectional comparisons, but cause less concerns for longitudinal comparisons. Also, considering students’ longitudinal improvements stratified by faculty participant (e.g., only looking within faculty participant) could alleviate some selection threats. For example, students are likely more homogeneous within a given class than across classes. There may be less confounding between population characteristics and effects of the ER intervention when looking only within one particular class (Shadish, Cook, & Campbell, 2002).

As depicted in Figure 9, it does not appear that maturation represented a substantial threat to internal validity. Recall, as discussed in Chapter Four, the effect of the ER intervention on students’ ER skills was not solely due to students’ general developmental level (i.e., freshmen, sophomore, junior, and senior). Lower-level students (e.g., freshmen) enrolled in a required general education course (i.e., Faculty #2) tended to earn higher ER-WR scores, on average, compared to upper-level students (e.g., juniors, seniors) enrolled in one of two elective major courses (i.e., Faculty #3 and #4). Yet, upper-level students enrolled in another elective course (i.e., Faculty #5) earned the highest ER-WR scores, on average. Therefore, observed improvements in students’ ER scores were likely not due to maturation effects.

Because the study was quasi-experimental there were several confounding variables that could not be controlled, and thus limited the study’s internal validity. These
variables included class size, class type (e.g., general education, elective, etc.), format in which class was taught (e.g., lecture, active learning, etc.), disciplinary context of the class, time of day class was taught, developmental level of student, and the teaching prowess of the faculty members instructing each of the six classes.

In comparison to the internal validity threats, there were relatively few external validity threats. External validity refers to the extent to which the study conclusions or inferences hold across varying students, settings, disciplines, institutions, etc. The extent to which study results can be generalized to varying contexts represents external validity (Shadish, Cook, & Campbell, 2002). Results suggested that the ER intervention positively influenced students across a variety of developmental levels and class types, meaning it may be generalizable across a variety of courses within different disciplines. Imagine that the intervention only had positive effects for the students in Faculty #2’s course (i.e., lower-level students in a required general education class). Then the conclusions or results may not be generalizable to upper-level students or students in elective classes. By including a variety of class types and students, some internal validity was sacrificed. However, external validity was strengthened because results suggested positive effects of the intervention, regardless of differences in students’ developmental stages, class types, etc. Also, I examined the ER intervention in the environments in which it would actually be occurring (i.e., live classrooms), which contributed to the ecological validity of the study. The setting and contexts of the study mirrored the real-world application of the intervention.

Additionally, the current study had several practical limitations. First, it is plausible that intervention features were implemented in the classrooms when the fidelity
researchers were not present to observe and record. Similarly, faculty participants could have implemented intervention features through email exchanges or conversations that took place outside of the classroom. For example, Faculty #4 reported that students occasionally discussed the 8KQ in their “reading notes” (i.e., notes that students took on assigned readings and were required to share with the faculty member) even though they were not explicitly instructed to do so. Additionally, Faculty #4 included an essay question related to the 8KQ on the final exam. Perhaps the fidelity data are not entirely complete; researchers likely missed the implementation of some amount of intervention features. Also, faculty could have implemented additional curricula or pedagogies that were not included as intervention specific features on the checklist. Supplemental curricula or pedagogies not explicitly captured by the checklist could have positively affected students’ ER skills. However, given they were not articulated in the checklist, they would not be represented in the fidelity data. Given these possibilities, the fidelity data are likely an underestimate of the extent to which students experienced ER education.

It is difficult to collect validity evidence to support interpretations of fidelity data. Observational data collection can introduce subjectivity or biases. And it is impossible to have a true “control” group in fidelity research. To address this potential study limitation, I used various procedures. Frist, as described in Chapter Three, I extensively trained researchers in implementation fidelity procedures, use of fidelity checklists, etc. Adequate training should help researchers apply the checklist in a consistent and accurate way. Second, I provided initial validity evidence for fidelity data by having multiple researchers collect fidelity data. Using two researchers should help combat subjectivity
and bias. Third, when only one researcher was available to collect data, I had faculty participants review those data for errors or omissions. Furthermore, fidelity data for each class session were adjudicated and averaged together. Adjudication processes promote consistency and accuracy, which are important precursors to providing validity evidence.

Further research is needed to build on this initial validity evidence. For example, future studies could recruit faculty who are experts in the MC and ER education. Such experts should know what quality implementation and high student responsiveness look like. Researchers could show the experts demonstrations or examples of what the ER intervention looks like when implemented with high versus low quality and/or with high versus low responsiveness. Providing such examples, and helping fidelity researchers internalize them, should promote more accurate fidelity data and more valid inferences.

Variations between faculty participants in perceived student responsiveness were an important component of the implementation fidelity results. But responsiveness is based on observers’ perceptions. To supplement these observations of responsiveness, faculty could ask students to autonomously self-report their own levels of responsiveness at the end of class. Perceived responsiveness gauged by observers could be averaged with students’ self-reported levels of responsiveness to obtain greater accuracy. Including student self-reported responsiveness data could provide further validity evidence for the inferences made from fidelity data.

**Future Research Directions.** I encourage MC leadership to continue partnering with faculty as they assess students’ ER skills, implement ER interventions, and re-assess to demonstrate learning improvement. Before collecting fidelity data for future studies, however, MC leadership and faculty members should consider revising the intervention
fidelity checklist. I have provided recommendations to guide checklist revisions. Additionally, if MC leadership and faculty decide to modify the checklist, I suggest retaining the original checklist for comparison studies. For example, MC leadership could conduct a study in which a subset of faculty participants implements the original checklist (with high fidelity), and another subset of faculty implements the modified version of the checklist (with high fidelity). Results could be used to determine whether students who received the original intervention demonstrated greater improvements compared to students who received the modified intervention. If students who received the modified intervention (e.g., a checklist including fewer features) demonstrated improvements similar to students who received the original intervention, then the modified intervention may be just as effective.

Higher education needs more examples of learning improvement studies. Given this study and Good’s (2015) study were both conducted at JMU, similar studies at different types of institutions (e.g., four-year, two-year, etc.) and programs (e.g., sciences, humanities, etc.) could illuminate generalizability questions. Perhaps MC leadership would consider partnering with universities that are interested in teaching the 8KQ framework, using the piloted intervention. Faculty from JMU could share assignments, lectures, and other resources. Or certain ethical reasoning classes could be taught simultaneously at multiple institutions. For instance, if several institutions wanted to implement the ER intervention in health science courses, they may be able to coordinate these efforts using existing telepresence systems.

The role of students’ cognitive development in their learning was beyond the scope of the current study. Therefore, future research could illuminate such
developmental levels, intervention effectiveness, intervention implementation, and improvements in students’ ER skills. Future studies should also emphasize the duration component of the fidelity checklist. Faculty participants were unable to articulate planned/intended duration times for the specific features or the more macro program components. MC leadership and faculty could use this study’s results to estimate intended duration times for the intervention specific features. Then future studies could take the next step of comparing actual and intended duration times.

As mentioned in Chapter Two, it may be difficult to sustain gains in students’ ER skills over time. Therefore, a logical next step would be to examine more long-term effects of the piloted intervention. Future research studies could assess a subset of participant students multiple semesters or years later (i.e., those who have not already graduated) to determine whether improvements in their ER skills were maintained. Additionally, if the current study were replicated, a third assessment data collection time point could be added. The third data collection time point could occur one to two years after students experienced the piloted intervention. Future research should also interview faculty participants to determine whether they are still implementing the piloted intervention even though the study has concluded. Perhaps faculty are continuing to use the Fidelity checklist and continuing to implement the intervention. Or perhaps they have abandoned the checklist and reverted back to previous curricula or pedagogies for teaching ER skills.

**Beyond JMU: Broader Implications for Higher Education**

Study results have implications for faculty, beyond JMU. Faculty from other institutions should consider including implementation fidelity in their teaching toolboxes.
Use of implementation fidelity checklists aids in aligning curricula and pedagogies with student learning outcomes and assessment instruments. Indeed, several faculty participants agreed that the implementation fidelity checklist positively affected their teaching and classroom culture. Faculty participants noted that use of the intervention fidelity checklist added structure to their teaching, kept them accountable, and allowed them to plan their courses with greater precision.

Results also have implications for institutions, beyond JMU. For instance, literature lacks examples of implementation fidelity research, as applied to academic programs. Through this study, I provided a detailed example of how to apply implementation fidelity research to a university-level academic program. Specifically, researchers demonstrated improved student learning and described how an educational intervention influenced learning improvement, using implementation fidelity data.

Furthermore, researchers provided an example of how implementation fidelity research, guided by the Simple Model (Fulcher et al., 2014), bridged the disconnect between learning outcomes assessment and learning improvement. Implementation fidelity data allow researchers to more closely align assessment results with educational interventions. Subsequently, other institutions could demonstrate improvements in learning and associate improvements with specific classroom experiences, through fidelity data.

As more institutions conduct implementation fidelity research and integrate fidelity data with assessment processes, practitioners will have greater evidence of their institution’s worth. Integrating assessment practices, like implementation fidelity, with curricula and pedagogies to demonstrate improvement can also support accreditation
efforts. For example, institutions accredited by SACSCOC are required to articulate, implement, and document continuous improvement initiatives.

More globally, questions of higher education’s worth and return on investment persist. Stakeholders must continuously evidence the value of higher education (Arum & Roksa, 2011; Taylor et al., 2011). Consider legislation from the Obama administration (e.g., college scorecard and gainful employment), which provided information on student loan debt, attendance costs, etc. for consumers (Lederman & Fain, 2017). Or the “Best Colleges” rankings produced by U.S. News & World Report. Colleges are rank-ordered based on a variety of metrics including faculty compensation, graduation rates, alumni giving, etc. (Rivard, 2014). However, there are alternative approaches for quantifying and reporting an institution’s value to students, lawmakers, and other consumers.

For instance, institutional value could be represented through evidence of learning improvement, and how well improvements are associated with learning interventions. What magnitudes of learning improvements can institutions empirically evidence? Can institutions empirically link learning improvements back to specific curricular and/or pedagogical experiences that faculty provided? Colleges struggle to answer such questions. Certainly, the majority of institutions have not been able to appropriately bridge the gap between assessment practices and educational interventions (i.e., curricula and pedagogies) (Banta, Jones, & Black, 2009; Blaich & Wise, 2011). But institutions could respond better by creating their own learning improvement examples. My hope is that this study provides a model for doing so.

As more institutions bridge the gap between assessment and learning, higher education will be one step closer to demonstrating worth. Worth that is not exclusively
proven via salary and loan repayment metrics, but through demonstrable learning improvement in areas that society desperately needs – like ethical reasoning.
Table 1

*Summary of Data Management and Motivation Filtering Process Results*

<table>
<thead>
<tr>
<th></th>
<th>ER-WR</th>
<th>ERIT</th>
<th>ER-WR</th>
<th>ERIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total students who attended data collection session</td>
<td>264</td>
<td>252</td>
<td>242</td>
<td>242</td>
</tr>
<tr>
<td>Total students whose data were actually recorded (i.e., total</td>
<td>260</td>
<td>252</td>
<td>240</td>
<td>242</td>
</tr>
<tr>
<td>number of students for which no Chromebook or data saving</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>issues occurred)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Of the students whose data were recorded, total students who</td>
<td>260</td>
<td>252</td>
<td>238</td>
<td>240</td>
</tr>
<tr>
<td>consented to participate in research project</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Of the students who consented to participate, total students</td>
<td>243</td>
<td>243</td>
<td>212</td>
<td>206</td>
</tr>
<tr>
<td>who passed effort subscore motivation filtering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of student essays deemed &quot;rateable&quot;</td>
<td>228</td>
<td>N/A</td>
<td>204</td>
<td>N/A</td>
</tr>
<tr>
<td>Total matched sample size for ER-WR and ERIT data sets</td>
<td>191</td>
<td>206</td>
<td>191</td>
<td>206</td>
</tr>
</tbody>
</table>
Table 2

Demographic Information for Students Included in Fall 2016 Pre/Post Matched ER-WR Sample, Fall 2016 Pre/Post Matched ERIT Sample, Fall 2015 Comparison ER-WR and ERIT Samples, and Spring 2016 Comparison ER-WR and ERIT Samples

<table>
<thead>
<tr>
<th>Age</th>
<th>ER-WR fall 2015</th>
<th>ER-WR spring 2016</th>
<th>ER-WR fall 2016*</th>
<th>ERIT fall 2015</th>
<th>ERIT spring 2016</th>
<th>ERIT fall 2016*</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>177</td>
<td>115</td>
<td>156</td>
<td>464</td>
<td>410</td>
<td>171</td>
</tr>
<tr>
<td>Female</td>
<td>60.45%</td>
<td>60.87%</td>
<td>84.62%</td>
<td>62.07%</td>
<td>61.95%</td>
<td>85.96%</td>
</tr>
<tr>
<td>Caucasian</td>
<td>80.79%</td>
<td>83.48%</td>
<td>89.10%</td>
<td>84.27%</td>
<td>85.85%</td>
<td>88.30%</td>
</tr>
<tr>
<td>Asian</td>
<td>9.04%</td>
<td>5.22%</td>
<td>4.49%</td>
<td>7.54%</td>
<td>5.37%</td>
<td>5.56%</td>
</tr>
<tr>
<td>African American</td>
<td>7.34%</td>
<td>6.09%</td>
<td>5.77%</td>
<td>7.33%</td>
<td>4.88%</td>
<td>5.85%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>5.65%</td>
<td>2.61%</td>
<td>1.92%</td>
<td>5.82%</td>
<td>5.85%</td>
<td>1.75%</td>
</tr>
<tr>
<td>Native American Indian</td>
<td>1.13%</td>
<td>0%</td>
<td>0%</td>
<td>0.86%</td>
<td>0.98%</td>
<td>0%</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>0.56%</td>
<td>0.87%</td>
<td>0.64%</td>
<td>0.86%</td>
<td>0.49%</td>
<td>0.58%</td>
</tr>
<tr>
<td>Not Specified</td>
<td>4.52%</td>
<td>4.35%</td>
<td>0.64%</td>
<td>3.66%</td>
<td>4.15%</td>
<td>1.17%</td>
</tr>
</tbody>
</table>

*Note. Means are presented with standard deviations in parentheses. Sample sizes for demographic information for some groups (e.g., fall 2015) are slightly smaller than sample sizes used in subsequent analyses because of the demographic information for some students were unavailable in assessment archival records. Also, certain demographic variables like SAT scores were excluded because students who participated in the current study did not necessarily consent for this information to be used as part of the current study. (*) denotes students who participated in current study and thus received some form of the piloted ER intervention. The demographic data for 35 of the student participants were not available in any archival assessment day files (e.g., due to students being transfer, etc.).
Table 3

**Average ER-WR Essay Scores by Element for Students Who Have Not Experienced Any Ethical Reasoning Interventions at JMU Compared to Students Who Have Experienced at Least One Ethical Reasoning Intervention at JMU**

<table>
<thead>
<tr>
<th>Rubric Element</th>
<th>fall 12’-spring 13’</th>
<th>fall 13’-spring 14’</th>
<th>fall 14’-spring 15’</th>
<th>fall 15’-spring 16’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 110 essays</td>
<td>N = 180 essays</td>
<td>N = 284 essays</td>
<td>N = 293 essays</td>
</tr>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A. Ethical Situation</td>
<td>1.56</td>
<td>0.90</td>
<td>1.94</td>
<td>1.16</td>
</tr>
<tr>
<td>B. Key Question Reference</td>
<td>0.76</td>
<td>0.58</td>
<td>1.13</td>
<td>0.94</td>
</tr>
<tr>
<td>C. Key Question Applicability</td>
<td>0.44</td>
<td>0.48</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>D. Ethical Reasoning: Analyzing individual KQ</td>
<td>0.48</td>
<td>0.54</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>E. Ethical Reasoning: Weighing the relevant factors &amp; deciding</td>
<td>0.50</td>
<td>0.55</td>
<td>0.90</td>
<td>0.83</td>
</tr>
<tr>
<td>OVERALL AVERAGE</td>
<td>0.75</td>
<td>0.61</td>
<td>1.13</td>
<td>0.79</td>
</tr>
</tbody>
</table>

*Note. SD = standard deviation and indicates the spread of scores around the mean. For example, a SD of about 1 on Element A with a mean of about 1 indicates that 68% of the essays (1 SD below and above the mean, assuming a normal distribution) received scores between 0 and 2. The scale is: 0 = Insufficient; 1 = Marginal; 2 = Good; 3 = Excellent; and 4 = Extraordinary.*
Table 4

*Initial Known Groups Validity Evidence for ER-WR Scores*

<table>
<thead>
<tr>
<th>Cohort #</th>
<th>Cohort Year</th>
<th>ER-WR Average Score</th>
<th>Comparison</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fall ‘12-spring ’13 (baseline- No ER intervention)</td>
<td>0.75</td>
<td>#1 vs #2</td>
<td>Students’ overall average ER-WR scores for FA13-SP14 were significantly higher than the overall average ER-WR scores for students assessed in FA12-SP13. On average, FA13-SP14 scores were about half of a standard deviation unit higher; ( t(288) = 4.32, p &lt; 0.0001 ), Cohen’s ( d = 0.52 ).</td>
</tr>
<tr>
<td>2</td>
<td>fall ‘13-spring ’14 (It’s Complicated ER intervention)</td>
<td>1.13</td>
<td>#1 vs #2</td>
<td>Students’ overall average ER-WR scores for FA14-SP15 were significantly higher than the overall average ER-WR scores for students assessed in FA12-SP13. On average, FA14-SP15 scores were about 0.37 of a standard deviation unit higher; ( t(392) = 3.14, p &lt; 0.001 ), Cohen’s ( d = 0.37 ).</td>
</tr>
<tr>
<td>3</td>
<td>fall ‘14-spring ’15 (It’s Complicated ER intervention)</td>
<td>1.05</td>
<td>#2 vs #3</td>
<td>Students’ overall average ER-WR scores for FA13-SP14 and FA14-SP15 were not significantly different; ( t(462) = 1.00, p = 0.159 ), Cohen’s ( d = 0.10 ).</td>
</tr>
<tr>
<td>4</td>
<td>fall ‘15-spring ’16 (It’s Complicated ER intervention)</td>
<td>1.39</td>
<td>#2 vs. #4</td>
<td>Students’ overall average ER-WR scores for FA15-SP16 were significantly higher than the overall average ER-WR scores for students assessed in FA12-SP13 (i.e., the baseline cohort). On average, FA15-SP16 scores were about one full standard deviation unit higher; ( t(401) = 9.16, p &lt; 0.0001 ), Cohen’s ( d = 1.02 ).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#1 vs. #4</td>
<td>Students’ overall average ER-WR scores for FA15-SP16 were significantly higher than the overall average ER-WR scores for students assessed in FA13-SP14. On average, FA15-SP16 scores were about 0.37 of a standard deviation unit higher; ( t(471) = 3.95, p &lt; 0.0001 ), Cohen’s ( d = 0.37 ).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#3 vs. #4</td>
<td>Students’ overall average ER-WR scores for FA15-SP16 were significantly higher than the overall average ER-WR scores for students assessed in FA14-SP15. On average, FA15-SP16 scores were about 0.45 of a standard deviation unit higher; ( t(575) = 5.39, p &lt; 0.0001 ), Cohen’s ( d = 0.45 ).</td>
</tr>
</tbody>
</table>
Table 5

*Variance Components, Relative Reliability, and Absolute Reliability Estimates for Pre- and Post-test ER-WR Scores for the (R:S)*E G-theory Design*

<table>
<thead>
<tr>
<th>Time</th>
<th>Description of Variance Component</th>
<th>Variance Component</th>
<th>Estimate</th>
<th>Time</th>
<th>Variance Component</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>Student (Object of Measurement)</td>
<td>$\sigma^2_S$</td>
<td>0.075</td>
<td>Post-test</td>
<td>$\sigma^2_S$</td>
<td>0.647</td>
</tr>
<tr>
<td></td>
<td>ER-WR rubric element</td>
<td>$\sigma^2_E$</td>
<td>0.292</td>
<td></td>
<td>$\sigma^2_E$</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>Rater nested within Student</td>
<td>$\sigma^2_{SR}$</td>
<td>0.242</td>
<td></td>
<td>$\sigma^2_{SR}$</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>Student*Element</td>
<td>$\sigma^2_{SE}$</td>
<td>0.031</td>
<td></td>
<td>$\sigma^2_{SE}$</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td>Student * Element * Rater</td>
<td>$\sigma^2_{SER,e}$</td>
<td>0.176</td>
<td></td>
<td>$\sigma^2_{SER,e}$</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>interaction, plus random error</td>
<td>$\sigma^2_{\delta}$</td>
<td>0.145</td>
<td></td>
<td>$\sigma^2_{\delta}$</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma^2_{\Delta}$</td>
<td>0.203</td>
<td></td>
<td>$\sigma^2_{\Delta}$</td>
<td>0.148</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G</td>
<td>0.340</td>
<td></td>
<td>G</td>
<td>0.841</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\phi$</td>
<td>0.269</td>
<td></td>
<td>$\phi$</td>
<td>0.814</td>
</tr>
</tbody>
</table>
Table 6

Variance Component for Object of Measurement, Relative Reliability, and Absolute Reliability Estimates for Pre-test ER-WR Scores Stratified by Rater Team for the (R*S*E)P G-theory Design

<table>
<thead>
<tr>
<th>Pair:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td># essays rated:</td>
<td>19</td>
<td>14</td>
<td>23</td>
<td>42</td>
<td>42</td>
<td>15</td>
<td>9</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>$\sigma^2_S$</td>
<td><strong>0.313</strong></td>
<td>0.046</td>
<td>0.065</td>
<td>0.063</td>
<td>0.030</td>
<td><strong>0.374</strong></td>
<td>0.000</td>
<td><strong>0.512</strong></td>
<td><strong>0.224</strong></td>
</tr>
<tr>
<td>G</td>
<td><strong>0.775</strong></td>
<td>0.338</td>
<td>0.576</td>
<td>0.429</td>
<td>0.256</td>
<td><strong>0.852</strong></td>
<td>0.000</td>
<td><strong>0.889</strong></td>
<td><strong>0.752</strong></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.639</td>
<td>0.236</td>
<td>0.330</td>
<td>0.238</td>
<td>0.106</td>
<td><strong>0.712</strong></td>
<td>0.000</td>
<td><strong>0.782</strong></td>
<td>0.568</td>
</tr>
<tr>
<td>Pre-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relSE</td>
<td>0.301</td>
<td>0.299</td>
<td>0.218</td>
<td>0.289</td>
<td>0.297</td>
<td>0.255</td>
<td>0.327</td>
<td>0.253</td>
<td>0.271</td>
</tr>
<tr>
<td>absSE</td>
<td>0.420</td>
<td>0.384</td>
<td>0.362</td>
<td>0.448</td>
<td>0.507</td>
<td>0.389</td>
<td>0.394</td>
<td>0.378</td>
<td>0.412</td>
</tr>
<tr>
<td>$\sigma^2_\delta$</td>
<td>0.091</td>
<td>0.089</td>
<td>0.047</td>
<td>0.084</td>
<td>0.088</td>
<td>0.065</td>
<td>0.107</td>
<td>0.064</td>
<td>0.074</td>
</tr>
<tr>
<td>$\sigma^2_\Delta$</td>
<td>0.177</td>
<td>0.148</td>
<td>0.131</td>
<td>0.201</td>
<td>0.257</td>
<td>0.152</td>
<td>0.155</td>
<td>0.143</td>
<td>0.170</td>
</tr>
</tbody>
</table>

| Weighted Average | G | 0.505 |  |  |  |  |  |  |  |
| | $\phi$ | 0.348 |  |  |  |  |  |  |  |

*Note. Rater pair #10 contained two raters who only rated one common essay; thus, Rater pair #10 could not be included in the g-theory analyses stratified by rater pairs. Thus, one essay could not be included in this analysis, bringing the total sample size for the G-theory analyses stratified by rater pairs to N = 190, instead of N = 191, which is the sample size for the (R:S)*E design results shown in Table 5.*
Table 7

*Rater Harshness ER-WR Essay Ratings for Implant (“Common”) Essay #1 Used During ER-WR Rating Session on Tuesday, January 3rd*

<table>
<thead>
<tr>
<th>RATER</th>
<th>RATER PAIRS</th>
<th>A. Ethical Situation</th>
<th>B. Key Question Reference</th>
<th>C. Key Question Applicability</th>
<th>D. ER: Analyzing</th>
<th>E. ER: Weighing and Deciding</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2.5</td>
<td>3.3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2.5</td>
<td>3.3</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2.5</td>
<td>3.4</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>3.5</td>
<td>3.6</td>
</tr>
<tr>
<td>H</td>
<td>4</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3.7</td>
</tr>
<tr>
<td>I</td>
<td>5</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>3.5</td>
<td>3.7</td>
</tr>
<tr>
<td>K</td>
<td>6</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3.7</td>
</tr>
<tr>
<td>L</td>
<td>6</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>3.5</td>
<td>3.7</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>3.8</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
<td>3.9</td>
</tr>
<tr>
<td>J</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.0</td>
</tr>
</tbody>
</table>
Table 8

*Rater Harshness ER-WR Essay Ratings for Implant (“Common”) Essay #2 Used During ER-WR Rating Session on Wednesday, January 4th*

<table>
<thead>
<tr>
<th>RATER</th>
<th>RATER PAIRS</th>
<th>A. Ethical Situation</th>
<th>B. Key Question Reference</th>
<th>C. Key Question Applicability</th>
<th>D. ER: Analyzing</th>
<th>E. ER: Weighing and Deciding</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2.8</td>
</tr>
<tr>
<td>L</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2.8</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2.5</td>
<td>2</td>
<td>2.9</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2.5</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>H</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>4</td>
<td>3.5</td>
<td>3</td>
<td>3</td>
<td>2.5</td>
<td>3.2</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2.5</td>
<td>3.3</td>
</tr>
<tr>
<td>I</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2.5</td>
<td>3</td>
<td>3.3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3.5</td>
<td>3.5</td>
<td>3</td>
<td>3.4</td>
</tr>
<tr>
<td>J</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3.4</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>2.5</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>3.5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3.5</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Table 9

Raw and Statistical Comparison of Students’ ER-WR Scores at Pre-test and Post-test for Each Rubric Element and for Overall (Average) ER-WR Scores

<table>
<thead>
<tr>
<th>Time</th>
<th>ER-WR Rubric Element</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A. Ethical Situation</td>
<td>2.260</td>
<td>0.607</td>
<td>0.000</td>
<td>3.500</td>
</tr>
<tr>
<td></td>
<td>B. Key Question Reference</td>
<td>1.067</td>
<td>0.596</td>
<td>0.000</td>
<td>3.500</td>
</tr>
<tr>
<td></td>
<td>C. Key Question Applicability</td>
<td>0.836</td>
<td>0.500</td>
<td>0.000</td>
<td>3.250</td>
</tr>
<tr>
<td></td>
<td>D. ER: Analyzing</td>
<td>0.831</td>
<td>0.543</td>
<td>0.000</td>
<td>2.750</td>
</tr>
<tr>
<td></td>
<td>E. ER: Weighing and Deciding</td>
<td>0.988</td>
<td>0.551</td>
<td>0.000</td>
<td>3.000</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>1.197</td>
<td>0.468</td>
<td>0.050</td>
<td>3.150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>ER-WR Rubric Element</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A. Ethical Situation</td>
<td>2.548</td>
<td>0.708</td>
<td>0.250</td>
<td>4.000</td>
</tr>
<tr>
<td></td>
<td>B. Key Question Reference</td>
<td>2.304</td>
<td>1.215</td>
<td>0.250</td>
<td>4.000</td>
</tr>
<tr>
<td></td>
<td>C. Key Question Applicability</td>
<td>1.928</td>
<td>1.169</td>
<td>0.000</td>
<td>4.000</td>
</tr>
<tr>
<td></td>
<td>D. ER: Analyzing</td>
<td>1.759</td>
<td>0.943</td>
<td>0.000</td>
<td>3.750</td>
</tr>
<tr>
<td></td>
<td>E. ER: Weighing and Deciding</td>
<td>1.577</td>
<td>0.915</td>
<td>0.000</td>
<td>3.750</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>2.023</td>
<td>0.877</td>
<td>0.200</td>
<td>3.850</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sig. Difference Pre to Post?</th>
<th>ER-WR Rubric Element</th>
<th>Mean Difference (Post-Pre)</th>
<th>SD of Difference</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A. Ethical Situation</td>
<td>0.288</td>
<td>0.828</td>
<td>4.81</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>B. Key Question Reference</td>
<td>1.237</td>
<td>1.223</td>
<td>13.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>C. Key Question Applicability</td>
<td>1.092</td>
<td>1.129</td>
<td>13.36</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>D. ER: Analyzing</td>
<td>0.928</td>
<td>0.951</td>
<td>13.48</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>E. ER: Weighing and Deciding</td>
<td>0.589</td>
<td>0.917</td>
<td>8.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>0.827</td>
<td>0.833</td>
<td>13.72</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Note. N = 191 for all matched pre/post ER-WR analyses; df = 190. Possible ER-WR overall scores range from 0-4.
Table 10

Cross-sectional Comparison of ER-WR Scores for Students Who Experienced It’s Complicated to Students Who Experienced the ER Intervention Piloted in the Current Study

<table>
<thead>
<tr>
<th>ER-WR Rubric Element</th>
<th>Mean (fall 2015)</th>
<th>Mean (fall 2016)</th>
<th>Mean (spring 2016)</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Ethical Situation</td>
<td>2.22</td>
<td>2.13</td>
<td>2.55</td>
<td>0.462</td>
</tr>
<tr>
<td>B. Key Question Reference</td>
<td>1.62</td>
<td>1.13</td>
<td>2.30</td>
<td>0.643</td>
</tr>
<tr>
<td>C. Key Question Applicability</td>
<td>1.32</td>
<td>0.87</td>
<td>1.93</td>
<td>0.611</td>
</tr>
<tr>
<td>D. ER: Analyzing</td>
<td>1.27</td>
<td>0.93</td>
<td>1.76</td>
<td>0.588</td>
</tr>
<tr>
<td>E. ER: Weighing and Deciding</td>
<td>1.14</td>
<td>0.98</td>
<td>1.58</td>
<td>0.538</td>
</tr>
<tr>
<td>Overall</td>
<td>1.51</td>
<td>1.21</td>
<td>2.02</td>
<td>0.659</td>
</tr>
</tbody>
</table>

Sig. Difference Between Non-Intervention and Intervention Groups?

<table>
<thead>
<tr>
<th>Rubric Element</th>
<th>Mean Difference (fall 2016-fall 2015)</th>
<th>t</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Ethical Situation</td>
<td>0.33</td>
<td>4.431</td>
<td>&lt; 0.001</td>
<td>0.462</td>
</tr>
<tr>
<td>B. Key Question Reference</td>
<td>0.68</td>
<td>6.170</td>
<td>&lt; 0.001</td>
<td>0.643</td>
</tr>
<tr>
<td>C. Key Question Applicability</td>
<td>0.61</td>
<td>5.874</td>
<td>&lt; 0.001</td>
<td>0.611</td>
</tr>
<tr>
<td>D. ER: Analyzing</td>
<td>0.49</td>
<td>5.647</td>
<td>&lt; 0.001</td>
<td>0.588</td>
</tr>
<tr>
<td>E. ER: Weighing and Deciding</td>
<td>0.44</td>
<td>5.168</td>
<td>&lt; 0.001</td>
<td>0.538</td>
</tr>
<tr>
<td>Overall</td>
<td>0.51</td>
<td>6.328</td>
<td>&lt; 0.001</td>
<td>0.659</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rubric Element</th>
<th>Mean Difference (fall 2016-spring 2016)</th>
<th>t</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Ethical Situation</td>
<td>0.42</td>
<td>4.959</td>
<td>&lt; 0.001</td>
<td>0.585</td>
</tr>
<tr>
<td>B. Key Question Reference</td>
<td>1.17</td>
<td>9.644</td>
<td>&lt; 0.001</td>
<td>1.138</td>
</tr>
<tr>
<td>C. Key Question Applicability</td>
<td>1.06</td>
<td>9.162</td>
<td>&lt; 0.001</td>
<td>1.081</td>
</tr>
<tr>
<td>D. ER: Analyzing</td>
<td>0.83</td>
<td>8.255</td>
<td>&lt; 0.001</td>
<td>0.974</td>
</tr>
<tr>
<td>E. ER: Weighing and Deciding</td>
<td>0.60</td>
<td>6.065</td>
<td>&lt; 0.001</td>
<td>0.716</td>
</tr>
<tr>
<td>Overall</td>
<td>0.81</td>
<td>8.817</td>
<td>&lt; 0.001</td>
<td>1.040</td>
</tr>
<tr>
<td>Time</td>
<td>Faculty</td>
<td># Students included in ER-WR Sample</td>
<td>ER-WR Rubric Element</td>
<td>Mean</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-----------------------------------</td>
<td>----------------------</td>
<td>------</td>
</tr>
<tr>
<td>Pre-test</td>
<td>Overall</td>
<td>194</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>77</td>
<td></td>
<td>A. Ethical Situation</td>
<td>2.214</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td></td>
<td>A. Ethical Situation</td>
<td>2.264</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td></td>
<td>A. Ethical Situation</td>
<td>2.500</td>
</tr>
<tr>
<td>Post-test</td>
<td>Overall</td>
<td>157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>77</td>
<td></td>
<td>A. Ethical Situation</td>
<td>2.357</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td></td>
<td>A. Ethical Situation</td>
<td>2.667</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td></td>
<td>A. Ethical Situation</td>
<td>2.179</td>
</tr>
</tbody>
</table>

**Table 11**

Longitudinal Comparison of ER-WR Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty
Table 12

Statistical Significance and Practical Significance for Longitudinal Comparison of ER-WR Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty

<table>
<thead>
<tr>
<th>Faculty #</th>
<th>Mean Difference (Post-Pre)</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.394</td>
<td>5.56</td>
<td>76</td>
<td>&lt;0.0001</td>
<td>0.824</td>
</tr>
<tr>
<td>2</td>
<td>1.161</td>
<td>5.82</td>
<td>17</td>
<td>&lt;0.0001</td>
<td>2.947</td>
</tr>
<tr>
<td>3</td>
<td>0.514</td>
<td>2.01</td>
<td>6</td>
<td>0.0916</td>
<td>1.702</td>
</tr>
<tr>
<td>4</td>
<td>1.214</td>
<td>4.49</td>
<td>6</td>
<td>0.0042</td>
<td>3.830</td>
</tr>
<tr>
<td>5</td>
<td>1.652</td>
<td>17.44</td>
<td>41</td>
<td>&lt;0.0001</td>
<td>2.838</td>
</tr>
<tr>
<td>6</td>
<td>0.630</td>
<td>5.47</td>
<td>39</td>
<td>&lt;0.0001</td>
<td>1.727</td>
</tr>
</tbody>
</table>
**Table 13**

*Longitudinal Comparison of ERIT Total Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty*

<table>
<thead>
<tr>
<th>Time</th>
<th>Faculty #</th>
<th># Students included in ERIT Sample</th>
<th>ERIT Total Score Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Time</th>
<th>Faculty #</th>
<th># Students included in ERIT Sample</th>
<th>ERIT Total Score Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-test</td>
<td>1</td>
<td>79</td>
<td>32.544</td>
<td>8.141</td>
<td>7.000</td>
<td>47.000</td>
<td></td>
<td>1</td>
<td>79</td>
<td>35.291</td>
<td>7.533</td>
<td>9.000</td>
<td>48.000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>18</td>
<td>35.667</td>
<td>5.698</td>
<td>24.000</td>
<td>45.000</td>
<td></td>
<td>2</td>
<td>18</td>
<td>38.833</td>
<td>5.533</td>
<td>24.000</td>
<td>47.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8</td>
<td>36.250</td>
<td>5.970</td>
<td>27.000</td>
<td>45.000</td>
<td></td>
<td>3</td>
<td>8</td>
<td>37.500</td>
<td>5.782</td>
<td>29.000</td>
<td>45.000</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9</td>
<td>33.111</td>
<td>9.597</td>
<td>11.000</td>
<td>46.000</td>
<td></td>
<td>4</td>
<td>9</td>
<td>38.444</td>
<td>7.485</td>
<td>20.000</td>
<td>45.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>45</td>
<td>34.733</td>
<td>7.165</td>
<td>16.000</td>
<td>49.000</td>
<td></td>
<td>5</td>
<td>45</td>
<td>37.778</td>
<td>6.399</td>
<td>18.000</td>
<td>48.000</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>47</td>
<td>31.745</td>
<td>6.716</td>
<td>21.000</td>
<td>45.000</td>
<td></td>
<td>6</td>
<td>47</td>
<td>34.234</td>
<td>7.322</td>
<td>19.000</td>
<td>46.000</td>
</tr>
</tbody>
</table>

*Note. Possible ERIT total scores range from 0-50.*
Table 14

<table>
<thead>
<tr>
<th>Faculty #</th>
<th>Mean Difference (Post-Pre)</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.747</td>
<td>4.60</td>
<td>78</td>
<td>&lt;0.0001</td>
<td>0.337</td>
</tr>
<tr>
<td>2</td>
<td>3.167</td>
<td>2.28</td>
<td>17</td>
<td>0.0355</td>
<td>0.556</td>
</tr>
<tr>
<td>3</td>
<td>1.250</td>
<td>0.99</td>
<td>7</td>
<td>0.3557</td>
<td>0.209</td>
</tr>
<tr>
<td>4</td>
<td>5.333</td>
<td>3.12</td>
<td>8</td>
<td>0.0142</td>
<td>0.556</td>
</tr>
<tr>
<td>5</td>
<td>3.044</td>
<td>4.04</td>
<td>44</td>
<td>0.0002</td>
<td>0.425</td>
</tr>
<tr>
<td>6</td>
<td>2.489</td>
<td>3.40</td>
<td>46</td>
<td>0.0014</td>
<td>0.371</td>
</tr>
</tbody>
</table>

Statistical Tests of Significance and Practical Significance for Longitudinal Comparison of ERIT Total Scores for Students Who Experienced ER Intervention Piloted in Current Study Stratified by Faculty
Table 15

Comparison of Current Study Results to Results from Good (2015)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A. Ethical Situation</td>
<td>2.36</td>
<td>0.87</td>
<td>1.94</td>
</tr>
<tr>
<td>B. Key Question Reference</td>
<td>1.44</td>
<td>0.92</td>
<td>1.13</td>
</tr>
<tr>
<td>C. Key Question Applicability</td>
<td>1.20</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td>D. ER: Analyzing</td>
<td>1.23</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>E. ER: Weighing and Deciding</td>
<td>1.15</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Overall</td>
<td><strong>1.47</strong></td>
<td><strong>0.74</strong></td>
<td><strong>1.13</strong></td>
</tr>
</tbody>
</table>
### Table 16

*Summary Information about Each of the Six Classes Included in the Current Study*

<table>
<thead>
<tr>
<th>Faculty #</th>
<th>Total number of students enrolled in course during fall 2016</th>
<th>Brief Description of Course</th>
<th>General Course Type</th>
<th>Total number of class sessions observed to collect fidelity data</th>
<th>Total number of class assignments used to collect implementation fidelity data</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>99</td>
<td>Upper level students; Required course for major; Ethics in class title</td>
<td>Lecture</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>#2</td>
<td>40</td>
<td>Lower level students; General Education Class; Fulfills Cluster 1 requirement; Ethics in class title</td>
<td>Lecture</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>#3</td>
<td>14</td>
<td>Upper level students; Elective Course</td>
<td>Seminar; Community Service Learning</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>#4</td>
<td>16</td>
<td>Upper level students; Elective Course</td>
<td>Seminar; Community Service Learning</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>#5</td>
<td>48</td>
<td>Upper level students; Course for minor</td>
<td>Lecture; Community Service Learning</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>#6</td>
<td>72</td>
<td>Upper level students; Required course for major</td>
<td>Lecture</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 17

*Implementation Fidelity Results for Duration, Student Responsiveness, Adherence, and Quality Aggregated Across All Faculty Members Stratified by MC ER Fidelity Checklist Overall Program Components*

<table>
<thead>
<tr>
<th>Overall Component of ER Piloted Intervention</th>
<th>Introduction/ Building Foundations to 8KQ</th>
<th>Case Study/ Dilemma Discussion</th>
<th>Examples</th>
<th>Visualization</th>
<th>Analysis w/ 8KQ</th>
<th>Weighing &amp; Deciding using 8KQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averages and Totals Across Faculty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>93</td>
<td>148</td>
<td>4</td>
<td>15</td>
<td>90</td>
<td>88</td>
</tr>
<tr>
<td>Student Responsiveness</td>
<td>4.14</td>
<td>4.44</td>
<td>4.75</td>
<td>4.31</td>
<td>4.37</td>
<td>4.58</td>
</tr>
<tr>
<td>Adherence</td>
<td>53.00</td>
<td>137.00</td>
<td>6.00</td>
<td>12.00</td>
<td>52.00</td>
<td>87.00</td>
</tr>
<tr>
<td>Quality</td>
<td>4.03</td>
<td>3.98</td>
<td>3.50</td>
<td>3.63</td>
<td>4.08</td>
<td>4.55</td>
</tr>
</tbody>
</table>

*Note. Adherence data are totals or frequency counts across all faculty members. Duration, Student responsiveness, and Quality data are averages across all faculty members. Duration data are in minutes.*
## Table 18

**Implementation Fidelity Results for Each Faculty Member Stratified by MC ER Fidelity Checklist Specific Features**

<table>
<thead>
<tr>
<th>Specific Features of ER Piloted Intervention</th>
<th>Introduction/Building Foundations to 8KQ</th>
<th>Case Study/Dilemma Discussion</th>
<th>Examples</th>
<th>Verification</th>
<th>Analysis of 8KQ</th>
<th>Weighing &amp; Deciding using 8KQ</th>
<th>Averages and Totals Within Faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student</strong></td>
<td><strong>Faculty #1</strong></td>
<td><strong>Faculty #2</strong></td>
<td><strong>Faculty #3</strong></td>
<td><strong>Faculty #4</strong></td>
<td><strong>Faculty #5</strong></td>
<td><strong>Faculty #6</strong></td>
<td><strong>Process</strong></td>
</tr>
<tr>
<td>Adherence</td>
<td>0.00</td>
<td>3.63</td>
<td>4.50</td>
<td>4.50</td>
<td>5.00</td>
<td>4.13</td>
<td>0.00</td>
</tr>
<tr>
<td>Quality</td>
<td>3.00</td>
<td>5.00</td>
<td>4.50</td>
<td>5.00</td>
<td>5.00</td>
<td>4.50</td>
<td>5.00</td>
</tr>
<tr>
<td><strong>Student</strong></td>
<td><strong>Faculty #1</strong></td>
<td><strong>Faculty #2</strong></td>
<td><strong>Faculty #3</strong></td>
<td><strong>Faculty #4</strong></td>
<td><strong>Faculty #5</strong></td>
<td><strong>Faculty #6</strong></td>
<td><strong>Process</strong></td>
</tr>
<tr>
<td>Adherence</td>
<td>0.00</td>
<td>4.00</td>
<td>4.50</td>
<td>5.00</td>
<td>5.00</td>
<td>4.13</td>
<td>0.00</td>
</tr>
<tr>
<td>Quality</td>
<td>4.50</td>
<td>4.00</td>
<td>4.50</td>
<td>4.50</td>
<td>4.50</td>
<td>4.50</td>
<td>4.00</td>
</tr>
</tbody>
</table>

**Legend:**
- **5.00**: Excellent
- **4.50**: Good
- **4.00**: Fair
- **3.50**: Poor

**Notes:**
- Faculty #1: [Details]
- Faculty #2: [Details]
- Faculty #3: [Details]
- Faculty #4: [Details]
- Faculty #5: [Details]
- Faculty #6: [Details]
Figure 1. Graphical Organizer of Study Terms

Note: the dashed lines in Figure 1 represent specific concepts and methods used as part of the study, to bridge the gap between assessment practices and demonstrable learning improvement.
Figure 2. Visualization and Description of the Steps in Fulcher and Colleagues (2014) Simple Model for Learning Improvement

**Assess**
- Direct measures
- Sound methodology
- Collect pre-intervention data

**Intervene**
- Identify 1 or 2 learning objectives to target
- Investigate current efforts
- Modify activities, curriculum, pedagogy, etc.
- Collect fidelity data

**Re-assess**
- Determine whether modifications positively influenced student learning
- Integrate fidelity and outcomes data to make informed revisions to interventions
Figure 3. Description and Visualization of the Process Used During the Summer Institute to Help Faculty Create an Ethical Reasoning Intervention and Fidelity Checklist
Figure 4. Depiction of the Two Facet Nested Design for G-theory Analysis with Shaded Black Area to Represent Error, Dotted Lines to Represent the Object of Measurement, and Solid Lines to Represent the Two Facets or Sources of Systematic Error
Figure 5. Comparison of Students from Good (2015) Study to Student Comparison Group Used in Current Study to Student Treatment Group Used in Current Study

<table>
<thead>
<tr>
<th>Group Description</th>
<th>Average ER-WR Essay Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (2015) Student in a control group who did not experience an ER infused course</td>
<td>0.00</td>
</tr>
<tr>
<td>Good (2015) Student in a treatment group who experienced an ER infused course</td>
<td>1.50</td>
</tr>
<tr>
<td>Students assessed during fall 2015 who did not receive the piloted ER intervention</td>
<td>2.00</td>
</tr>
<tr>
<td>Students assessed during spring 2016 who did not receive the piloted ER intervention</td>
<td>1.00</td>
</tr>
<tr>
<td>Students assessed during fall 2016 who did receive the piloted ER intervention from the current study</td>
<td>3.50</td>
</tr>
</tbody>
</table>
Figure 6. Number of Total Intervention Specific Features Implemented by Each Faculty Member Plotted in Comparison with Their Students’ Average Post-test ER-WR Essay Scores
Figure 7. Students’ Average Post-Test ER-WR Total Scores Plotted Against Average Perceived Student Responsiveness and Quality Ratings for All Faculty Participants with Red Line Representing University Strategic Plan Goal for 2020.
*Figure 8.* Faculty Participant Profiles of Students’ Average Post-Test ER-WR Total Scores, Average Perceived Student Responsiveness, and Quality Ratings with Red Line Representing University Strategic Plan Goal for 2020
Figure 9. Faculty Participant Profiles of Student ER-WR Average Scores, Average Perceived Student Responsiveness, and Average Quality for All Faculty Participants Categorized According to Class Type (Orange = General Education Class with Lower Level Students, Grey = Primarily Active Learning or Seminar Based, Non-Required Course with Service Learning Component and Upper Level Students, Purple = Primarily Lecture Based, Required Course for Major with Upper Level Students) with Red Line Representing University Strategic Plan Goal for 2020
Figure 10. Faculty Participant Profiles of Student Z-score ER-WR Average Scores, Z-score Average Perceived Student Responsiveness, Z-score Average Quality, and Z-score Adherence
## Appendix A

Implementation Fidelity Checklist Used to Gather Fidelity Data for JMU Orientation Program

<table>
<thead>
<tr>
<th>Transfer Orientation</th>
<th>2015 Implementation Fidelity Checklist</th>
<th>Auditor:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Objective

**Academic Requirements Knowledge:** Upon completion of Transfer Summer Springboard, students will be able to correctly identify the academic requirements for major, degree, and graduation completion at JMU.

<table>
<thead>
<tr>
<th>Program Component</th>
<th>Duration</th>
<th>Comments/Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Planned</td>
<td>Actual</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specific Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 credit on transcript intended for major, go to major department head RKM5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adherence</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No</td>
<td>1 = Low (confusing)</td>
</tr>
<tr>
<td></td>
<td>3 = Medium</td>
</tr>
<tr>
<td></td>
<td>5 = High (clear)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General Education Presentation 2 (Arin, Curt, Catlyn)</th>
<th>25 min. (50 min. total)</th>
<th>The minimum GPA required to graduate is 2.0 RKM8</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 credit on transcript intended for major, go to major department head RKM5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>000 credit on transcript intended for Cluster 1, go to General Education RKM4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General Education Presentation 1 (Meg)</th>
<th>25 min. (50 min. total)</th>
<th>Honor Code needs to be completed as soon as possible RKM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 credit on transcript intended for Cluster 1, go to General Education RKM4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

212
<table>
<thead>
<tr>
<th>Objective</th>
<th>Program Component</th>
<th>Duration</th>
<th>Comments/Responsiveness</th>
<th>Specific Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Planned</td>
<td>Actual</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>N/A</td>
<td>n/a</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Academic Requirements Knowledge Continued</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Modules (Completed before coming to TSB)**
  - [http://www.jmu.edu/transfers-1st-semester](http://www.jmu.edu/transfers-1st-semester)
  - Send final transcript to Admissions[^RKM3]
  - Minimum number of credit hours to graduate (120)[^ARK1]
  - Minimum number of credit hours from a 4-year institution to graduate (60; 50%)[^ARK2/3]
  - Minimum number of credits from JMU to graduate (30; 25%)[^ARK4]
  - AP and IB credits must be sent directly to JMU Admissions[^ARK6]

[^RKM3]: Minimum number of credit hours to graduate (120) [ARK1]
[^ARK1]: Minimum number of credit hours from a 4-year institution to graduate (60; 50%) [ARK2/3]
[^ARK2/3]: Minimum number of credits from JMU to graduate (30; 25%) [ARK4]
[^ARK4]: AP and IB credits must be sent directly to JMU Admissions [ARK6]
<table>
<thead>
<tr>
<th>Objective</th>
<th>Program Component</th>
<th>Duration</th>
<th>Planned</th>
<th>Actual</th>
<th>Comments/Responsiveness</th>
<th>Specific Features</th>
<th>Adherence</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 = Low (unengaged)</td>
<td></td>
<td>Yes/No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 = Medium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 = High (engaged)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Acclimation and Community Building: As a result of attending Transfer Summer Springboard, students will significant increase in their cohesion to the JMU community</td>
<td>Transfer Advising</td>
<td>20-60 min</td>
<td>NA</td>
<td>NA</td>
<td>Overview of the major</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Description of the major admission and progression standards</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Explanation of the recommended and required courses for your major</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Provide an overview of the first semester major course selection</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Welcome</td>
<td>35 min.</td>
<td></td>
<td></td>
<td>TOPA cheer!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peer Discussion</td>
<td>60 min.</td>
<td></td>
<td></td>
<td>Sarah’s welcome and intro by university administration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>President Alger’s inspirational speech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Faculty academic speech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All TOPA introduction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Icebreakers/ Name games</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Post-it Note Activity: 1. Excited, 2. Concerned, 3. Questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ceremoniously receive JAC</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Objective</td>
<td>Program Component</td>
<td>Duration</td>
<td>Planned</td>
<td>Actual</td>
<td>Comments/Responsiveness</td>
<td>Specific Features</td>
<td>Adherence</td>
<td>Quality</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------</td>
<td>----------</td>
<td>---------</td>
<td>--------</td>
<td>-------------------------</td>
<td>-------------------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Getting JAC&lt;sup&gt;PCS&lt;/sup&gt;</td>
<td>2 min.</td>
<td></td>
<td>N/A</td>
<td>1 = Low (unengaged)</td>
<td>Dining</td>
<td>Yes/No</td>
<td>Meal plans are purchased at Card Services RKM1</td>
</tr>
<tr>
<td></td>
<td>Student Services Video</td>
<td>28 min.</td>
<td></td>
<td></td>
<td>3 = Medium</td>
<td></td>
<td></td>
<td>Tuition payments are made through the Business Office RKM2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 = High (engaged)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Resource Knowledge:** As a result of attending Transfer Summer Springboard, students will demonstrate an increase in knowledge of JMU resources by correctly matching resources and how they address student needs.
<table>
<thead>
<tr>
<th>Objective</th>
<th>Program Component</th>
<th>Duration</th>
<th>Comments/Responsiveness</th>
<th>Specific Features</th>
<th>Adherence</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Planned</td>
<td>Actual</td>
<td></td>
<td>Yes/No</td>
<td>1 = Low (confusing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sarah reviews the checklist that students have completed after the video</td>
<td></td>
<td>3 = Medium</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Business Office[^RKM2]</td>
<td></td>
<td>5 = High (clear)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Admissions[^RKM3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Off-Campus Life</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Health Center</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Information Technology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Appendix B

**Performance Assessment Rubric Used to Rate Student ER-WR Essays**

<table>
<thead>
<tr>
<th>Insufficient 0</th>
<th>Marginal 1</th>
<th>Good 2</th>
<th>Excellent 3</th>
<th>Extraordinary 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>No reference to decision option(s).</td>
<td>Implicit reference to decision options AND/OR little context given regarding decision option(s).</td>
<td>Explicit but unorganized reference to decision option(s) and context.</td>
<td>Clear description of decision option(s) and context.</td>
<td>Meets criteria for Excellent AND…</td>
</tr>
<tr>
<td>Reference to zero or only one key question.</td>
<td>Vague references to key questions OR only two key questions referenced.</td>
<td>References four key questions.</td>
<td>References six key questions.</td>
<td>References all eight key questions.</td>
</tr>
<tr>
<td>No rationale provided for the applicability or inapplicability of any KQs to the ethical situation.</td>
<td>Provides a rationale for the applicability or inapplicability of two key questions to the ethical situation.</td>
<td>Provides a rationale for the applicability or inapplicability of four key questions to the ethical situation.</td>
<td>Provides a rationale for the applicability or inapplicability of six key questions to the ethical situation.</td>
<td>For all eight questions provides a rationale for its applicability or inapplicability to the ethical situation.</td>
</tr>
<tr>
<td>No attempt to analyze any of the referenced key questions.</td>
<td>Analysis attempted using two or more key questions. Typically incorrect ascription of the key questions to the ethical situation. Account is unclear, disorganized, or inaccurate.</td>
<td>Analysis attempted using three or more key questions. Basically accurate ascription of the key questions to the ethical situation. Account is unclear or disorganized.</td>
<td>Analysis attempted using three or more key questions. Accurate ascription of the key questions to the ethical situation. Account is clear and organized.</td>
<td>Meets criteria for Excellent AND…</td>
</tr>
<tr>
<td>No judgment is presented OR judgment presented with no rationale.</td>
<td>Uses products of the analysis and provides some weighing to make a decision. Account is unclear, disorganized, or inaccurate.</td>
<td>Conveys weighing approach using analysis products. Provides an intelligible basis for judgment.</td>
<td>Meets criteria for Good AND…. Logically terminates in decision that will be reached.</td>
<td>Meets criteria for Excellent AND…</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Products of analysis weighed to make judgment compelling.</td>
</tr>
</tbody>
</table>

- Context treated with nuance
- Builds tension with organization and word choice.
Appendix C

Writing Prompt Provided to Students for ER-WR Essay Assessment

Often in life, we encounter situations that are ethically complicated. For example, if you saw a hungry child steal fruit from a grocery store, you’d likely think of many reasons to report the person and many reasons not to do so. The faculty and staff at JMU are interested in the ethical reasoning thought process in which students engage when confronted with such situations.

For this assessment, please…

(1) Explain a complicated, ethically significant choice you faced: a choice that required a lot of thinking and deliberation.

(2) Indicate the ethical considerations that you deem relevant to this particular situation and why, as well as which ones are not relevant and why.

(3) Be sure to clarify your ethical reasoning process as much as possible. Try to provide an ethical analysis that is as rich and multifaceted as possible.

(4) Lastly, be sure to say what decision you made and why.

You will have 60 minutes to compose this essay. Your document should contain no fewer than 250 words.

Please feel free to express whatever opinions you might hold. Your essay will NOT be evaluated on what decision was made, but rather the clarity and complexity of the thought process underlying that decision.

Thank you!
## Appendix D

### Faculty Development Summer Institute Week Schedule At-a-Glance

<table>
<thead>
<tr>
<th>Day 1 ➔ Implementation Fidelity Basics</th>
<th>Activities/Curriculum</th>
<th>Objective Covered</th>
</tr>
</thead>
</table>
| **Intro to Learning Improvement Model:** Understanding and Applying Implementation Fidelity; Linking these two processes together | - Brief intro to the research project- explaining why we are all here; the need for this research project; review faculty signed MOU’s and faculty responsibilities/roles in the project  
- Brief intro to assessment cycle  
- Brief intro to the Simple Model for learning improvement  
- Introduce implementation fidelity through examples from JMU’s campus and introduce very general idea of backward design  
- Discuss the five components of Implementation Fidelity  
  - Think. Pair. Share- Work with partner to fill in a blank implementation fidelity checklist for one intervention that you do in your class (can pick any intervention/activity/assignment, etc.)  
  - What was the hardest part about creating the checklist? What components require further clarification?  
  - Explain how implementation fidelity information can be useful pedagogically and useful for demonstrating learning improvement  
- Describe the typical Implementation Fidelity data collection process  
  - JMU’s Orientation Program  
  - JMU’s LID CIS project  
- Group discussion about the implementation fidelity matrix of possible inferences (Gerstner & Finney, 2013)  
  - Work through four (hypothetical) examples set in an academic contexts using the fidelity matrix (Gerstner & Finney, 2013) to convey the importance of fidelity data when making inferences based on outcomes assessment data | - Describe the steps of the assessment cycle  
- Explain how assessment practice and teaching and learning are connected or related  
- Identify and describe the steps in the Simple Model  
- Identify the five components of implementation fidelity  
- Explain the steps or process of collecting implementation fidelity data  
- Articulate why implementation fidelity data is important for demonstrating student learning improvement  
- Create a “general” implementation fidelity checklist aligned with the ER intervention and MC SLOs 4 & 5 | Kristen Smith |

Day 1 Wrap Up: tie back to why we are here: to apply implementation fidelity principles to Ethical Reasoning Instruction and to give faculty members development opportunities and skills that they can use beyond this research project. Tomorrow we will review the MC objectives and discuss ER interventions

<table>
<thead>
<tr>
<th>Days 2, 3, &amp; 4 ➔</th>
<th>Activities/Curriculum</th>
<th>Objective Covered</th>
<th>Facilitators/Co-facilitators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Brief review the “program differentiation” component of implementation fidelity</td>
<td>- Discuss and agree upon key</td>
<td>Kristen Smith</td>
</tr>
<tr>
<td>Application of Implementation Fidelity to ER Education</td>
<td>Creating an ER Intervention &amp; accompanying fidelity checklist mapped to MC SLOs 4 &amp; 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Brief review of MC 8 Key Questions</td>
<td>• Review the MC cognitive SLOs &amp; the current institution-wide interventions that are mapped to each (e.g., It’s Complicated, also cover the MCI intervention)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Review the MC cognitive SLOs &amp; the current institution-wide interventions that are mapped to each (e.g., It’s Complicated, also cover the MCI intervention)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o Majority of focus on <strong>SLO 4 &amp; 5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Think. Create. Pair. Share- Individually, articulate the key features of what you believe would be a “highly effective” MC ER reasoning intervention aligned with SLO 4 &amp; 5 that you could do in your classroom. Discuss in small groups and as larger group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o In order for students to be able to do SLO 4 &amp; 5, what do we need to have them practice in our classrooms? What general things or “key features” must students do in order to achieve SLO 4 &amp; 5? How can these be generalized across disciplines? How can I teach students these things or integrate these “key features” into my course?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Integrate these key features into a clear, agreed upon list of key intervention features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o General “Key features” must be agreed upon by all faculty participants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Provide “blank” fidelity checklist and have faculty fill in with agreed upon key features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o this will be the final checklist used for data collection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>components or features of an effective ER intervention aligned with MC SLOs 4 &amp; 5</td>
<td>• Based on those agreed upon components, design an ER intervention aligned with MC SLOs 4 &amp; 5 that can be applied in various classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Create a “general” implementation fidelity checklist aligned with the ER intervention and MC SLOs 4 &amp; 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Day 5</th>
<th>Finalizing ER intervention, checklist, &amp; scheduling class observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Faculty complete filling in fidelity checklist with agreed upon key features</td>
<td>• Review implementation fidelity data collection procedures for Fall 2016</td>
</tr>
<tr>
<td>• Review implementation fidelity data collection procedures for Fall 2016</td>
<td>• discuss expectations for faculty “self-audit” using the fidelity checklist</td>
</tr>
<tr>
<td>• discuss expectations for faculty “self-audit” using the fidelity checklist</td>
<td>• Create schedule for when Kristen will observe classes to collect implementation fidelity data</td>
</tr>
</tbody>
</table>

Smi th, Sara Finn ey, Bill Ha wk, Lori Pyle
Appendix E

Faculty Memorandum of Understanding (MOU) for Participating in Madison Collaborative Implementation Fidelity Research Project

I, ____________________ _______________________, agree to be a faculty participant in the “Aligning Assessment and Pedagogy via Implementation Fidelity” research project starting in Summer 2016 through the end of the Fall 2016 semester. As a faculty participant in this research project, I agree to fully complete each of the following roles, responsibilities, and conditions.

I hereby agree to:

- attend and participate in a week-long implementation fidelity and ER intervention workshop to be held during the summer of 2016, during which I will collaborate with other faculty to develop an ethical reasoning intervention (e.g., activities, assignments, demonstrations, etc.)
- commit to teaching ER using the 8KQ in at least one of my classes during the Fall 2016 semester.
- dedicate at least ten hours or more of class time to an MC ER intervention/instruction. Note, these ten hours can be inside and/or outside of class meeting times.
- allow co-principle investigators to observe at least 6 of my class sessions and collect implementation fidelity data.
- fill out the implementation fidelity checklist for myself (i.e., “self-audit”) for at least three class sessions
- allow co-principle investigators to collect outcomes assessment data from my students at both PRE and POST-test (e.g., the first week and the end of the semester, respectively). Note, the co-principle investigators can collect outcomes assessment data outside of regularly scheduled class meetings, and via the Ashby assessment lab, if that is more convenient for the faculty member.
- work with the co-principle investigators to create a schedule of classroom observation dates/times for the Fall 2016 semester.
- receive payment of a stipend of $2,000 from the MC for my participation in the entire duration of this research project (e.g., all training, classroom observations, etc. occurring during Summer 2016 and Fall 2016).
# Appendix F

**Madison Collaborative Ethical Reasoning Intervention Fidelity Checklist**

Fidelity Researcher: ___________________________ Date of Data Collection: ______________

<table>
<thead>
<tr>
<th>Overall OBJ</th>
<th>Specific Program OBJ</th>
<th>Program Component</th>
<th>Duration in min. (Actual)</th>
<th>Responsiveness</th>
<th>Adherence</th>
<th>Quality</th>
<th>Specific Features</th>
<th>Comments/Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC SLOs 4 &amp; 5</td>
<td>A, B, C, D</td>
<td>Introduction / Building Foundation to 8KQ</td>
<td></td>
<td>1 = Low (unengaged) 3 = Medium 5 = High (engaged)</td>
<td></td>
<td></td>
<td>Elaborate or unpack each of the 8KQ (e.g., reviewing the handbook, lecturing, PPTs, video clip, discussion,)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Read/Review SLOs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Read/Review rubric</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Students experience a “check point” to check their own knowledge of the 8KQ (maybe use Bill’s “ERIT” items??; crossword puzzle or word find; ball activity, news stories)</td>
<td></td>
</tr>
<tr>
<td>A, B, C, D, E</td>
<td>Case Study (Dilemma Discussion)</td>
<td>Map 8KQ to some other work (can be something disciplinary like standards or something societal like policies or media or something practical, or something personal, news stories, onto class community or rules of engagement, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A, B, C, D, E</td>
<td>Case Study (Dilemma Discussion)</td>
<td>Critique/edit/comment/annotate the 8KQ (e.g., could be wiki, could be collectively done in class, what do you like about 8KQ? What would you change about them?; collective knowledge building)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A, B, C, D, E</td>
<td>Case Study (Dilemma Discussion)</td>
<td>Provide/discuss/present example of a decision making process with AND without ethical reasoning (“ethical reasoning” is defined as being able to use 2+ KQ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Review/Refresh 8 KQ

Identify where/how each of the 8KQ are/ are not applied within the case

Give/discuss rationale for how each of the 8KQ are/are not applied

Engage in reflection (e.g., could be formal or informal, written, oral, group, what issues did you have, what was easy/hard)
<table>
<thead>
<tr>
<th>A, B, C, D, E</th>
<th>Examples</th>
<th>Identify/discuss which (if any) aspects of the case are “compelling?” To what extent or degree was the case “compelling?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, C, D, E</td>
<td>Multi-modal Analysis Visualization</td>
<td>Have students together review/build a “strong” or “effective” example of ethical reasoning (e.g., show Bill students’ videos in class and talk about what they could have done differently)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify and explain how characteristics or features make the case (in)effective referencing SLOs and/or rubric?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Students experience (either visually or through some other sensory modality like touch, feel, movement, etc.) analysis processes- this can be “shown” by professor or created by students (e.g., block exercise, using color or size, show Keston PPT slide, students personify KQ using their bodies as visuals, concept map- decision trees, Pictionary type game, role playing, collages, etc.)</td>
</tr>
<tr>
<td>A, B, C, D, E</td>
<td>Analysis of/with KQ</td>
<td>Students experience some sort of analysis (or breaking a part) of at least one KQ; should get at nuances if possible</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify obstacles or pitfalls to analysis (e.g., only analyzing 1 KQ, confirmation bias, privilege)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consider contextual factors (e.g., could include or &quot;get at&quot; multiple perspectives)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expose/demonstrate/suggest how multiple perspectives can compete/interact w/one another within the same KQ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A, B, C, D, E</th>
<th>Weighing &amp; Deciding using 8KQ as rationale</th>
<th>Students process something (debate, case, discussion, etc.) using 8KQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Students must arrive at or grapple with a particular conclusion or decision point</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple stakeholders and/or multiple perspectives are identified or considered</td>
</tr>
</tbody>
</table>
Appendix G

Student Opinion Survey (SOS)

Please think about the test or tests that you just completed. Mark the answer that best represents how you feel about statements 1 through 10 below.

A= Strongly Disagree
B=Disagree
C=Neutral
D=Agree
E=Strongly Agree

1. Doing well on these tests was important to me.
2. I engaged in good effort throughout these tests.
3. I am not curious about how I did on these tests relative to others.
4. I am not concerned about the scores I receive on these tests.
5. These were important tests to me.
6. I gave my best effort on these tests.
7. While taking these examinations, I could have worked harder on them.
8. I would like to know how well I did on these tests.
9. I did not give these tests my full attention while completing them.
10. While taking these tests, I was able to persist to completion of the tasks.
References


