Spring 2015

Improving student learning in higher education: A mixed methods study

Megan R. Good
James Madison University

Improving Student Learning in Higher Education: A Mixed Methods Study

Megan Rodgers Good

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May, 2015

Acknowledgements

I am fortunate to have many colleagues who supported me throughout my doctoral journey. First, I want to sincerely thank my mentor, Dr. Keston Fulcher. I am so grateful for your support and encouragement. You taught me to "riff" and "suspend all disbelief"--- skills that will undoubtedly serve me throughout my career.

I substantially benefitted from my terrific dissertation committee, all of whom contributed unique perspectives that were invaluable. Dr. Cara Meixner, my dissertation co-chair, introduced me to the world of qualitative research and faculty development. Thank you for broadening my horizon. I am also grateful to my other committee members— Drs. Jeanne Horst, John Hathcoat, and Lori Pyle. Thank you for making my dissertation stronger and supporting me throughout the entire process.

I will look back on my graduate years with great fondness, largely because of the many relationships I formed with my fellow graduate students. Specifically, I would like to thank Jerusha Gerstner, Bo Bashkov, Kristen Smith, Chris Coleman, Daniel Jurich, Jason Kopp and Becca Marsh Runyon. I am so thankful for the support network you provided and I'm grateful to know each of you as a colleague and lifelong friend.

My dissertation would have been impossible without the support of the Madison Collaborative. Bill Hawk and Lori Pyle were supportive of this work from the start and I am I so thankful. Also, I would like to thank Melissa Grant, Rebecca Jones, Rebecca Redman, Hannah McEwen, and Mikala Morrow. Not only did you help me collect data, you also served as a breath of fresh air and constantly reminded me of the value that students bring to bear in any conversation. I would also like to give special thanks to Liz

Hawk Sanchez, who helped me become a better writer and made me more confident in my writing ability along the way. You're a winner Liz, thank you!

Finally, I'd like to thank my family. Your support and pride means everything to me. To my wonderful husband AJ-- I appreciate your willingness to deliver pizza to participants, but not as much as I appreciate the constant source of support and love you provided me during the hardest parts of this journey. You make me a stronger person and I am deeply grateful to have you in my life.

Table of Contents

List of Tables

List of Figures

Abstract

To improve quality, higher education must be able to demonstrate learning improvement. To do so, academic degree program leaders must assess learning, intervene, and then re-assess to determine if the intervention was indeed an improvement (Fulcher, Good, Coleman, and Smith, 2014). This seemingly "simple model" is rarely enacted in higher education (Blaich & Wise, 2011). The purpose of this embedded mixed methods study was to investigate the effectiveness and experience of a faculty development program focused on a specific programmatic learning outcome. Specifically, the intervention was intended to increase students' ethical reasoning skills aligned with a university-wide program. The results suggested that this experience did indeed improve student's ethical reasoning skills. Likewise, the experience was positive for faculty participants. This study provides evidence supporting the connection of assessment and faculty development to improve student learning.

CHAPTER 1

Introduction

The United States was once regarded as a global leader in higher education (Flannery, 2011). Surprisingly, our nation now places 12th worldwide (The White House, n.d.a). This fact, along with a general concern about the value of college, primes higher education to improve. A commitment to improvement would produce stronger graduates who learned more than their predecessors. Likewise, such a commitment would benefit the United States' position in global rankings.

Three areas challenge higher education and need improvement: cost, access, and quality (Reindl, 2007). That is, the *cost* of higher education is rising which makes *access* more difficult for students of lower economic means. *Quality* refers to the quality of education provided to students and ultimately their success after graduation. At the heart of the quality concern is the question, "How well are students doing [in terms of learning]?" (Reindl, 2007, p. 3).

Many current efforts are addressing the cost and access concerns. For example, President Obama's College Scorecard makes an institution's value and affordability transparent to prospective students and their families (The White House, n.d.b). Likewise, The Bill and Melinda Gates Foundation and The Lumina Foundation have dedicated resources to improve student access to higher education (Bill and Melinda Gates Foundation, n.d.; Lumina Foundation, 2013). While the federal government and private foundations are tackling access and affordability challenges, concerns of quality are the responsibility of regional accreditation. According to the United States Department of Education's website, "The goal of accreditation is to ensure that education provided by

higher education institutions meets acceptable levels of quality" (U.S. Department of Education, 2013).

Accreditation

Six regional accrediting agencies oversee a peer-review process of accreditation: Higher Learning Commission (HLC; 2014), Middle States Commission on Higher Education (MSCHE; 2011), Northwest Commission on Colleges and University (NWCCU; 2010), New England Association of Schools and Colleges- Commission on Institutions of Higher Education (NEASC-CIHE; 2011), Southern Association of Colleges and Schools- Commission on Colleges (SACSCOC, 2012), and WASC Senior College and University Commission (WASC-SCUC; 2013). Accreditation encourages institutions to create goals "for self-improvement of weaker programs and (stimulate) a general raising of standards among educational institutions" (U.S. Department of Education, 2013, "Accreditation in the U.S."). Further, institutions must be accredited in order for their students to receive federal financial aid (U.S. Department of Education, n.d.).

To operationalize quality, each regional accreditor uses a set of evaluation criteria (e.g., graduation rates, fiscal responsibility, the role of governing boards). A primary measure of quality in higher education is student learning. Therefore, each accreditor has at least one specific criterion pertaining to student learning. Specifically, all regional accreditors require evidence of student learning *improvement* based on assessment results (HLC, 2014; MSCHE, 2011; NEASC-CIHE, 2011; NWCCU, 2010; SACSCOC, 2012; WASC-SCUC, 2013). For example, SACSCOC criteria 3.1.1.1 states, "The institution identifies expected outcomes, assesses the extent to which it achieves these outcomes,

and provides *evidence of improvement* [emphasis added] based on analysis of the results in each of the following areas: …educational programs, to include student learning outcomes…(SACSCOC, 2012, p. 27)."

Thus, regional accreditors expect academic programs to assess student learning and use the results to evidence improvement. Although this expectation for quality is clear, very few examples of programmatic learning improvement exist (Blaich & Wise, 2011). In fact, Kuh and Ewell (2010) stated that using assessment data for learning improvement remains "the most important unaddressed challenge related to student learning outcomes assessment in our country" (p. 24).

From Assessment to Learning Improvement

To meet accreditation standards, programs must assess student learning and use the assessment results to improve their programs. It is surprising that so few institutions can evidence learning improvement given that most are assessing student learning (Kuh and Ikenberry, 2009). One possible explanation is that there are ample resources available on assessment mechanics (e.g., Banta, Lund, Black & Oblander, 1996; Suskie, 2010), but until recently, scarce information has been available on *how* to use results.

Fulcher, Good, Coleman, & Smith (2014) provided a framework on how to evidence learning improvement: assess, intervene, re-assess. As a first step, a program must have a robust assessment process (i.e., the assessment mechanics must be strong). Following, a program must do something differently, or intervene. This might include adding courses, changing course sequences, or adjusting pedagogies. Finally, after new cohorts of students experience the revised curriculum, a program re-assesses student learning to determine if the programmatic "intervention" was in fact an improvement.

As previously stated, there are many resources available on best practices in assessment. Nevertheless, there are few resources that focus on using assessment results. Fulcher et. al (2014) are among the first to discuss how a program might "intervene" by providing a four step process to improve student learning. Programs begin by selecting a target objective to improve, then faculty explore current program efforts, next the faculty propose learning modifications (e.g., changes to the curriculum, courses, sequencing), and finally, the program faculty produce an improvement timetable.

Fulcher et al. (2014) note that this is a multiyear process because programs must wait for new cohorts of students to experience the revised curriculum. In their article, the authors recommended that assessment practitioners partner with faculty development experts on campus to support programs in these endeavors. However, the authors do not explain what such a partnership would entail.

Statement of the Problem

In the United States, higher education needs improved student learning to enhance overall quality. Currently, regional accreditors evaluate quality by requiring institutions to evidence learning improvement; however, examples of learning improvement are rare (Blaich & Wise, 2011). Fulcher et al. (2014) provided a conceptual framework on *how* to evidence learning improvement and suggested connecting assessment with faculty development, although, the authors did not detail what this partnership would look like.

In this dissertation I demonstrate how learning improvement can be achieved by integrating efforts from assessment and faculty development offices. In addition, this dissertation is among the first to provide a scholarly research contribution to evidencing learning improvement. Thus, the purpose of this embedded mixed methods study is to

investigate the effectiveness and experience of a faculty development program focused on a specific programmatic learning outcome. The faculty development program, or faculty "learning intervention," is intended to improve student learning at the program level.

CHAPTER 2

Literature Review

Purpose of the Literature Review

The purpose of this literature review is to provide a rationale for integrating assessment with faculty development programming to improve student learning. I begin by discussing how the aforementioned use of results problem is a validity issue. Next, I anchor this study in the assessment literature, where authors have noted the learning improvement problem. Because the use of assessment results, or "closing the loop," is rare, I include extant literature on the topic. Following, I explore the faculty development domain, where this dissertation's learning intervention is situated. Finally, I discuss ethical reasoning, the learning outcome of interest in this study.

Construct Validity and Systemic Validity

In educational testing, validity is "…the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (AERA, APA, & NCME, 1999, p. 9). This understanding of validity is often referred to as construct validity. According to Benson (1998), a construct "represents an abstract variable derived from observation or theory" (p. 10). Thus, construct validation is a process by which test scores acquire meaning (Benson, 1998).

Threats to validity are many times defined in terms of construct under-representation and construct-irrelevant variance (Messick, 1995). Construct underrepresentation occurs when a measurement instrument is too narrow and does not include important facets of the construct of interest, while construct-irrelevant variance arises when an instrument measures facets outside of the construct (Messick, 1995).

Not all researchers limit their conceptualization of validity to test score inferences. For example, Frederiksen and Collins (1989) explained that when a test is introduced into a dynamic educational system, the test can change the system by affecting instruction. Instructional changes caused by the existence of a test are the basis for "systemic validity" (p.27). A researcher would have evidence of systemic validity if he could demonstrate that a test caused changes in learning.

Frederiksen and Collins (1989) explained how traditional multiple-choice tests can threaten systemic validity using a geometry test example that required students to perform a geometric proof. In this case, there were only 12 possible proofs a student could be asked to perform. Rather than teaching students mathematical reasoning skills, students were taught to memorize the 12 proofs that could be on the test. Thus, the educational system was compromised by the introduction of the high-stakes geometry test.

A researcher pondering the validity of test scores would say the geometry example is a classic case of construct irrelevancy. However, Frederiksen and Collins (1989) argued it is also a violation of systemic validity because the test changed how students were being taught. The example is problematic because students received limited instruction on geometrical reasoning; rather, memorization skills were reinforced because of the narrow nature of the high stakes test.

As a solution, Frederiksen and Collins (1989) recommended performance assessments yield better systemic validity. The idea is that teachers are going to teach to the test; however, this can be a positive outcome when a test measures higher order skills of interest. The authors directed attention to the assessment system by stating, "The goal

of assessment has to be, above all, to support the improvement of learning and teaching" (p. 32).

Like Frederikson and Collins, education and measurement leaders have oft cited performance assessment as being authentic. Nevertheless, performance assessments pose unique validity challenges. To this issue, Messick (1994) reviewed three authors' perspectives on performance assessment validity including Frederiksen and Collins' (1989) article. Messick did not wholly agree with Frederiksen and Collins's view, pointing out that they wrongly assumed all other aspects of the educational system were working. Messick believed that this notion of validity is too limited stating, "…the issue is not just the systemic validity of the tests but rather the validity of the system as a whole for improving teaching and learning (p.16)." Messick drew attention to the validity of the system rather than just a test's effect on a system.

This notion of systemic validity is similar to internal validity, which evaluates the cause-and-effect relationship between variables (Barron, Brown, Egan, Gesualdi, & Marchuk, 2008). As demonstrated by the regional accreditation standards, assessment in higher education is intended to be a catalyst for learning improvement. Said another way, assessment should cause learning improvement. However, if intended changes never occur, one may argue that the current systems are not valid in terms of internal validity or the systemic validity terms discussed by Messick (1994). Thus, underuse of assessment results is a threat to the validity of assessment practice. To understand assessment and its relationship to learning improvement, one must begin by exploring assessment mechanics.

Assessment of Student Learning

The goal of student learning outcomes assessment is to gather reliable data about student learning that can be used to improve an academic program if undesirable results are discovered (Erwin, 1991; Pepin, 2014). Since the initial 1980s call for assessment in higher education, institutions have slowly initiated assessment processes, although these processes vary in quality (Kuh et al., 2014).

Traditional Assessment Process. Program level (e.g., Biology, BA) assessment of student learning outcomes begins with programs clearly defining objectives by describing the knowledge, skills, and abilities expected of graduates. Once objectives are defined and articulated in student-centered, measurable, and specific terms, faculty must then provide a conceptual guide to where learning is thought to occur in the curriculum; a process called curriculum mapping, which aligns student learning outcomes with required courses and experiences in their curriculum (Palomba & Banta, 1999).

Following objective-curriculum alignment, the program must either create or select measurement instruments that align with stated objectives before employing an appropriate research design to collect data. Regarding data collection, the program should consider issues of sampling and student motivation. Representative sampling is necessary to make generalizable assertions about student learning. For example, sampling only honors students would lead to a biased representation of programmatic learning. Likewise, students should put forth their best performance on the assessment instrument; if they do not, perhaps because they lack motivation, they may rush through the test or answer randomly, yielding lower test scores (Wise & DeMars, 2005).

After data have been collected, reliability and other psychometric properties should be estimated. Next, the program must interpret the assessment results in reference to the program objectives. The hope is that programs will use this information to make systematic changes to the curriculum based on the assessment results (Erwin, 1991). This process is known as "closing the loop" (Banta & Blaich, 2011, p. 22).

Unfortunately, programs can get stuck trying to perfect the assessment mechanics. An assessment coordinator may not be satisfied with a certain measure or data collection design and spend time changing assessment details. A perfect assessment process does not exist, however. While it is necessary for a program to have trustworthy data, at some point the program must realize that their process is "good enough" and begin a conversation about curricular and/or pedagogical changes (Blaich & Wise, 2010). If a program strives for perfect assessment, student learning will never be affected.

Assessment as a Process of Inquiry. Jonson, Guetterman, and Thompson (2014) suggested approaching assessment as a process of inquiry, likening assessment to traditional research initiatives. In this framework, the first step focuses on a student learning question of interest. Second, faculty members gather student data that addresses their question. Third, the faculty members interpret and evaluate the data by "engaging stakeholders in *meaning making*" (p.19); a process involving open dialogue. Lastly, the fourth step is to use assessment results for improved teaching and learning.

Although unstated by Jonson et al. (2014), the mechanics of the proposed approach are the same as the traditional approach (e.g., there are design considerations); the only difference is in the way in which assessment is described. In Jonson et al.'s framework there is more intentional time spent discussing assessment results.

Unfortunately, like the traditional approach, the leap between making sense of assessment data and the use of results is undefined. In fact, Jonson and colleagues (2014) acknowledged that "…intention does not guarantee that an improvement will occur, and often whether the learning improvement does occur is not determined" (p. 24). In their paper, Jonson et al. investigated the use of results by evaluating assessment reports at a single institution using a program evaluation framework. However, Jonson and her colleagues found only a few cases (21%) of assessment reports with evidence of instrumental use or changes to a program based on assessment results.

Closing the Loop

Leading assessment authors almost always include "use of results" or "closing the loop" as a step in the assessment process (e.g., Huba & Freed, 2000; Walvoord, 2004). Discussions of the use of results are typically embedded within assessment books and are brief. Recently, Fulcher et al. (2014) provided a model for unpacking the term "use of results." The authors noted that some practitioners defined use of results as changes made to assessment mechanics or programmatic changes (e.g., changes to curricula or pedagogy). However, they pointed out that a change is not an improvement. Rather, "A change is only an improvement when one can demonstrate its positive effect on student learning" (Fulcher, et al., 2014, p. 4). Fulcher et al. (2014) stated that in order to evidence learning improvement, a program must assess, intervene, and then re-assess. Only when re-assessment reveals greater learning proficiency can a program state that learning improvement occurred. Figure 1 provides a visual depiction of Fulcher et al.'s (2014) simple model for learning improvement.

The notion of assess-intervene- re-assess looks like a traditional pre-post assessment design. A traditional pre-post assessment tests the same set of students at two different time points. However, it should be noted that in the simple model assessment is of graduating *cohorts.* Thus, the model is not a within-subjects design, but rather between-subjects (i.e., different cohorts of students).

Suskie (2010) referred to "closing the loop" in the same way as Fulcher et al, (2014) describing a hypothetical situation where faculty members deemed students' writing scores weak based on assessment results. To address this issue, the faculty members introduced a new problem-based learning strategy across the curriculum. The next time they assessed student writing, the instructors discovered that writing scores had improved. Again, a program must *re-assess*; a pedagogical change alone is necessary, but is not sufficient, for program improvement.

Frequency of Learning Improvement: Rare

The simple model (Fulcher et al., 2014) boils down learning improvement to its most basic form, although the general idea is not new. Blaich and Wise (2011) investigated the frequency of such evidenced improvement by investigating the national Wabash study findings. The Wabash College Center of Inquiry developed a national longitudinal assessment study designed to "deepen our understanding of the teaching practices, student experiences, and institutional conditions that promote the development of students' critical thinking, moral reasoning, leadership towards social justice, well-being, interest in engagement with diversity and interest in deep intellectual work" (Blaich & Wise, 2011, p.7). Many volunteer institutions participated in this longitudinal study where students were assessed when they entered college, after their first year, and

again during their senior year (Center of Inquiry: Wabash College, 2009). Participating institutions adopted a common battery of quality instruments and Wabash Study researchers provided detailed assessment reports on the behalf of these institutions. Thus, participation in the Wabash Study yielded high quality assessment results for institutions to consider.

When designing the study, the researchers operated on three core assumptions about what helps and hinders effective assessment. First, they believed that a lack of high quality data impeded institutions from using assessment results to improve student learning. Second, the researchers posited that detailed reports of assessment data would initiate campus conversations about improving student learning. Finally, the researchers assumed that the intellectual approach that faculty use to engage with their scholarship would facilitate the creation of assessment projects that improved student learning (Blaich & Wise, 2011).

From the beginning, researchers emphasized gathering and analyzing *quality data*. The researchers did not consider asking institutions what they would do after assessment reports were in hand. To their dismay, institutions overwhelmingly did not use their assessment results to improve student learning. Blaich and Wise (2011) concluded that their three core assumptions about closing the loop were wrong; they reflected that too much time was spent focusing on the data and too little time was spent on using assessment results. The researchers postulated two major reasons for the stagnation: first, data collection easily became a routine; there was little attention paid to reviewing the data, and second, Blaich and Wise assumed that the data would "speak loudly enough" to warrant action (p.12). However, this assumption does not consider the

many demands on faculty members' time; faculty have teaching, scholarship, and service responsibilities that often take precedence to interesting assessment findings that warrant additional research (which equals additional time). Blaich and Wise also noted that it is far less risky to continually analyze data instead of acting on assessment results.

Blaich and Wise's (2011) article is a critical contribution to the assessment domain. Most work on assessment focuses on methods; this is one of the first articles to draw attention to the use of results dilemma. In this piece, Blaich and Wise reported quantitative assessment results from the Wabash Study and reflected on their surprising finding: institutions rarely used assessment results. Unfortunately, the authors did not follow up with institutions to ask why assessment data were not used. Nevertheless, Blaich and Wise's reflections advised leaders in the field to address this issue.

Soon after Blaich and Wise's paper was published, the editor of *Change* magazine asked Trudy Banta, a leader in the assessment field, and Charles Blaich to co-write an article about how institutions have used their assessment results. After thoroughly reviewing the literature and relying on personal consultation experiences, Banta and Blaich concluded that such examples were extremely rare (2011).

Instead of writing about how assessment results are used, Banta and Blaich (2011) focused on factors they perceived to impede assessment data use. Like Blaich and Wise (2011), Banta and Blaich (2011) stated that gathering and analyzing data are not enough to demonstrate improved learning. They conceded "… even the most beautifully collected and interpreted evidence will have no impact on students whatsoever unless it engages an institution's faculty, staff, governance structures, faculty development programs, and leaders" (Banta & Blaich, 2011, p.23).

Banta and Blaich (2011) provided recommendations for solving the use issue: first, faculty members and other leaders on campus should engage with the assessment process. Second, external mandates should facilitate campus engagement with assessment, although they currently do not do so. Third, Banta and Blaich recommended using local measures or connecting measures to individual courses to facilitate understanding of assessment results. Finally, the authors described turnover as an impediment to use of assessment results. When faculty members, administrators, or assessment practitioners leave a position, predecessors will likely have different views on the assessment process and using results, thus sending different messages.

Additionally, Banta and Blaich (2011) recommended that institutions regularly evaluate whether or not 1) they are providing adequate resources for faculty to use results, 2) they are communicating the results effectively, and 3) student learning data is reaching potential users. Of note, the authors recommended that institutions should spend more time and money on *using* assessment results rather than on *gathering them*. They stated, "If all of an assessment program's resources are gobbled up gathering evidence, no change is likely to occur" (p. 26). Programs must invest in the activities that support programmatic changes based on assessment data.

Blaich and Wise (2011) successfully captured the state of affairs: everyone is doing assessment and almost no one is using the results to improve student learning. Banta and Blaich (2011) also recognized this issue and even provided a few insights about why institutions have trouble using results. Unfortunately, the authors of both articles did not provide a concrete solution or "how to" guide for practitioners.

Case Studies. Albeit rare, some institutions are closing the loop. *Change* magazine highlighted Kaplan University. The National Institute on Learning Outcomes Assessment (NILOA) provided several other examples.

Reed, Levin, and Malandra (2011) described the process that the for-profit institution, Kaplan University, used to close the assessment loop. In 2008, Kaplan transitioned from a program-level portfolio assessment system to a "course-level assessment system" (p.45) to create a tighter feedback loop for using assessment results to make curricular and pedagogical changes. At Kaplan, it was difficult for faculty members to translate program assessment results to their specific classes. By moving the level of assessment to individual courses, faculty could more readily understand and use the results.

The course-level assessment system is multi-tiered: each academic degree program has learning outcomes and each course within the program has course objectives that are tied to program outcomes. Specific assignments are aligned with each outcome and rubrics with a common rating scale (1-5) are used to assess the outcomes. Kaplan created teams of faculty, subject-matter experts, and curriculum-design experts to create the outcomes, assignments, and rubrics for each course. Faculty members at Kaplan use a common database to enter their classroom assessment results. The database system can generate feedback to individual faculty and create reports for administrators. It is important to note that the performance of students is *not* used to evaluate the performance of faculty members (Reed et al., 2011).

The Kaplan database system also has a mechanism to record when faculty members make changes to their courses based on assessment data and subsequently these

systems can generate pre-post intervention comparison reports. For example, Kaplan faculty members found a course, with many sections, that had low assessment scores regarding students' ability to design a website. The department head and faculty members decided to try four different course changes and measure the efficacy of each. This process not only allowed for an experimental approach to identify what worked, but it also gave the faculty an opportunity to engage in the Scholarship of Teaching and Learning (SoTL). Subsequently, faculty members at Kaplan documented student growth and were able to identify the most effective change enacted. Thus, Kaplan University had evidence of closing the loop.

According to Reed et al. (2011), Kaplan University owes its success to a few key factors, including the availability of sufficient resources, champions of the cause from across the university, and planning for data usage from the beginning. Because executives were on board with the assessment strategy, academic leadership was empowered to drive the changes, faculty members were engaged in the project, and institutional research and faculty development staff provided support. Thus, all members of the university were champions of the cause. However, it was not clear how the faculty development staff were involved.

Reed et al (2011) also shared lessons learned along the way: institutions need a "safe haven" (p.52) for discussing the tensions between having a centralized curriculum and respecting individual teaching styles, and, discipline is required in order to stay focused on the information needed to make changes. They explicitly avoided the temptation to collect more information just "because [they] can" (p.52). Reed et al. (2011) ended the article with future goals, including the investigation of the psychometric

properties of their instruments and developing feedback loops of different lengths (e.g., immediate employment for students).

Kaplan University created a unique assessment system with results-use in mind from the start. The authors defined assessment appropriately (i.e., assess-intervene-re-assess) and used teams of knowledgeable constituents to create a centralized curriculum. They also were among the few institutions with an example of improved learning based on assessment results.

Although Kaplan University engaged in an outstanding process, there are a few areas of concern. First, it must be recognized that there is value in program level assessment; it provides a snapshot of cumulative growth and development at the end of students' coursework. By only focusing on the course level, it is difficult to make inferences about what students know, think, or can do at the *end* of their program. Additionally, it appears that reliability and validity information is unknown. Quality assessment should precede use of results so that changes are based on trustworthy data. Finally, non-profit universities cannot exert the same level of curricular control as for-profit institutions such as Kaplan. Faculty from non-profit institutions would theoretically disagree with a standardized curriculum arguing it infringes upon their academic freedom (Hara, 2010).

Following Blaich and Wise's (2011) article, NILOA researchers (Baker, Jankowski, Provezis, & Kinzie, 2012) sought to identify and learn more about institutions that use assessment results to improve student learning. To do so, NILOA researchers conducted nine case studies across institutions practicing high quality assessment. After considering institution type and geographic location, they used the following selection

criteria: institutions recommended by NILOA's National Advisory Panel member nominations, the Council for Higher Education Accreditation (CHEA) award winners, and institutions from NILOA's prior research regarding assessment practice in the field (Baker et al., 2012). This approach yielded the selection of nine very different institutions.

Once institutions were selected, the researchers inquired about the individuals on campus most familiar with assessment (e.g., director of assessment), requested their participation in an interview, and conducted 60-minute phone interviews. Each researcher also thoroughly read material on the institution's website. After the nine case studies were completed, the researchers collectively discussed themes among the cases and reported a summary of common themes in a single NILOA white paper (Baker et al., 2012).

Across the nine case studies, Baker et al. (2012) found that the institutions sampled were universally focused on using assessment results to improve student learning. Likewise, all case study site constituents felt that they still had room to improve and that they had not yet "arrived" (p.6) at their ultimate goal. All case study sites were working to advance four common areas: focusing assessment efforts, harnessing accountability for internal improvement, communicating widely about assessment, and allowing time for internal stakeholders to make meaning of and to reflect on assessment results. To supplement Baker and colleagues' (2012) integration and summary, NILOA posted the nine case studies on their website, which provided more depth on assessment and results-use at each of the nine institutions. Upon careful review of each study, three

case studies, summarized below, stood out as exemplars: Capella University, St. Olaf College, and Carnegie Mellon University.

Capella University is a fully online for-profit institution, with tightly integrated curriculum, pedagogy, and assessment (Jankowski, 2011). In fact, collaborative teams of faculty, curriculum specialists, instructional designers, course developers, and assessment specialists work together to embed assessment throughout the institution using a four-phase backward design process: define, design, develop, and deliver. Define refers to the creation of outcomes; that is, what is it that students should know, think, or do as a result of the program? The design phase includes curriculum and instructional designers creating the course. Because Capella is fully online, the development phase includes the course integration into a learning management system. Lastly, instructors review and deliver the course in the final stage (Jankowski, 2011).

Capella reports assessment data to its governing board and is held accountable for improving the student experience. The tight alignment of curriculum, pedagogy, and assessment makes it obvious where learning interventions should occur. This alignment, coupled with administrative support, facilitates true loop closure.

Capella has two examples of using results, one in business and one in psychology. In both examples, undesirable results were found for a specific learning outcome. The department heads critically examined the alignment maps and course activities within their respective programs. Each department head identified areas that could be strengthened. After working with faculty to make changes, both programs saw growth (Jankowski, 2011).

Capella University is similar to Kaplan University in several ways. Both are for-profit online institutions with strong curricular control. Unlike Kaplan, however, Capella's assessment efforts are focused at the program level. Both institutions designed their curricula with teams of experts and had continuous improvement in mind from the start.

In contrast, St. Ola College is a small, private, liberal arts college in Minnestoa (Jankowski, 2012). The institution operates on a five-year assessment cycle, where every fourth year is deemed a "reflection year" and no data are collected. One faculty member described the Assessment Director as "more of a coach than anything else" (Jankowski, 2012, p. 3). The Assessment Director regularly works with the faculty development office on campus, which emphasizes the Scholarship of Teaching and Learning. Together, these offices helped to create a culture of systematic inquiry into student learning.

St. Olaf has a use-focused approach to assessment that permeates the academic culture. The environment at St. Olaf is very supportive of scholarly inquiry into student learning. However, within this case study, there was no evidence of true loop closure, although the college does have many unique structures (Jankowski, 2012). The use-focused approach and partnership with the faculty development office seem to have greatly contributed to faculty buy-in on campus. Also, the embedded reflection time is unique and prevents the institution from routinely collecting data as an exercise, as Blaich and Wise (2011) warned against.

Carnegie Mellon University (CMU) is a private research university in Pittsburgh, PA (Kinzie, 2012). One of the hallmarks of CMU that advances its program level

assessment and efforts to improve student learning is that CMU has institutionalized data-driven decision making by integrating assessment into the academic program review process (called the President's Advisory Board, or PAB). Of note, the President and upper administrators were involved with the recommendation process of the PAB. Assessment is a key data source in this process, which encourages continuous improvement on campus.

Another CMU hallmark is the Eberly Center for Teaching Excellence. Not only does it promote evidence-based practice from the learning sciences literature, it is also the hub for assessment activities. Thus, assessment support is housed in the faculty development center, which strengthens the relationship of and communication between assessment, teaching, and learning.

At CMU, there is one reported example of true loop closure. In the engineering department, assessment results highlighted that students lacked the experimental knowledge that the faculty expected. The NILOA report did not specify the definition of experimental knowledge or how it was deficient. Nevertheless, faculty members collectively agreed to teach experimental knowledge in two new courses. This endeavor was supported by the Eberly Center, which provided workshops and consultations to help make curricular changes that ultimately led to an increase in student learning. CMU has a unique structure that incorporates assessment for improvement into university leadership conversations. Likewise, CMU has housed assessment support within its faculty development office, a place already seen as a resource for teaching and learning.

In higher education, institutions are rarely able to evidence improved learning after reflecting upon assessment results. There are a few exceptions, however. The two

for-profit institutions, Capella University and Kaplan University, created centralized program curricula with a tight alignment between objectives, course activities, and assessment. This alignment facilitates the use of results by pinpointing areas in need of improvement. In non-profit institutions, which are typically less flexible, this systematic approach is rare and often impractical.

In these examples, a few key factors contributed to success. At St. Olaf College, built-in reflection time for assessment helped to change academic culture. At CMU, a key feature was assessment's role in the academic governance structure. Both approaches are novel. Likewise, many of the case studies pointed to faculty development offices as being an important partner in learning improvement initiatives.

Faculty Development

Original forms of faculty development in higher education supported faculty members' pursuit to stay abreast in their field. However, in the 1970s, the higher education landscape began to change when the baby-boomers flooded colleges. This surge in enrollment coupled with student protests about "irrelevant courses and uninspired teaching" spurred faculty development efforts with the focus on enhancing teaching (Gaff & Simpson, 1994, p.168). By the 1980s, a new series of academic challenges arose; specifically, there was a call for increased quality and coherence in general education and majors. Thus, faculty development became the vehicle for guiding curricular changes. Once a new curriculum was approved, faculty developers hosted seminars and workshops on content, course design, and innovative instructional techniques (Gaff & Simpson, 1994).

Faculty development centers are now thriving on many campuses. Purposes vary, but most centers aim to improve teaching and learning (Lee, 2010). Steinert et al. (2006) sought to gauge the effectiveness of such programs. These researchers systematically reviewed the faculty development literature for medical teachers, focusing on faculty development programs aimed to improve teaching effectiveness. The criteria for inclusion in Steinert and colleagues' review were the following: studies with a faculty development focus, studies targeting basic science and clinical faculty members in medicine, and studies that measured program effectiveness beyond satisfaction. All reviewed articles were published between 1980 and 2002 in English, French, Spanish, or German. The duration of the reviewed faculty development programs ranged from a 1.5 hour workshop to a semester.

Steinert et al.'s (2006) review of the effectiveness of such programs suggests that faculty development programs are beneficial. Specifically, medical faculty members reported benefits such as satisfaction, positive changes in attitudes toward teaching, increased knowledge of education principles, gains in teaching skills, and positive changes in teaching behavior. Only three of the studies investigated by Steinert et al. measured impact beyond the individual faculty member.

Faculty Development and Student Learning. Steinert et al. (2006) identified three studies (out of 53) that were focused on "change among the participants' students, residents, or colleagues" (p.501). One of the three studies, conducted by Nathan and Smith (1992), evaluated the impact of teacher-training workshops on student evaluations of teaching for 12 medical faculty members by measuring the difference in evaluation scores before and after the training. On average, student ratings of teaching increased

significantly following the workshop. However, the researchers did not find statistically significant differences in student learning, measured by student exam scores. Details about the measures used in this study were not provided.

In the second study Marvel (1990) attempted to improve faculty clinical teaching skills by providing feedback in the same way students are provided feedback (i.e., parallel process). Ten faculty members participated. These 10 faculty members were videotaped during teaching sessions and evaluated on seven teaching skills prior to the intervention using a behaviorally anchored rubric. The researcher installed a video camera into the classroom weeks prior to the study to desensitize faculty members to its presence.

After recording baseline data, Marvel (1990) scheduled a 45-minute feedback session with each of the 10 faculty members. During this session, the video recording was shown and the faculty member was given the opportunity to evaluate herself using the behavioral rubric, which constituted the feedback session. The author then re-assessed the faculty member after five teaching sessions to determine if teaching skills had improved. Additionally, the researcher gathered resident (i.e., student) perceptions of the faculty members' teaching abilities before and after the intervention. The researcher also gathered patient ratings of residents to determine if the faculty's improved teaching techniques impacted their residents' interviewing skills, thereby increasing patients' perception of the residents.

On average, five of the seven teaching behaviors increased after the feedback intervention. Residents rated faculty members high at both time points, although there were gains on only two of the seven teaching skill areas. Patient ratings of residents were

slightly higher in five of the seven teaching areas, but these results were not statistically significant.

Marvel's (1990) study provided some evidence of a faculty development experience's impact on teaching skills. Unfortunately, it is unclear from this study if students learned more as a result. Students rated the faculty members higher in certain areas, but it is unknown if student *learning* and skills were improved.

Skeff, Stratos, Campbell, Cooke, and Jones (1986) were the last group to be identified in Steinert et al.'s (2006) study as measuring change beyond the faculty member. Unlike the previous two studies, Skeff et al. (1986) used an experimental design. Like Marvel (1990), the purpose was to improve faculty members' teaching skills. Forty-six faculty members were assigned to either a control group (i.e., no intervention) or a seminar training session with other faculty members. The researchers evaluated teaching before and after the intervention using four measures: videotaping, teaching evaluations completed by students, faculty questionnaire about their teaching, and a student questionnaire about the impact of the faculty member on the student's knowledge, skills, and attitudes. Levels of all dependent variables increased in the experimental group but not for the control group. Student learning was not investigated.

The three studies described above demonstrate evidence of the effectiveness of faculty development programs on faculty members' teaching skills. Likewise, they suggest that students feel positively about their professor's teaching changes. However, the ultimate dependent variable – student learning – was neglected in each design.

Steinert et al.'s (2006) review occurred within the medical field and the articles of interest took place over twenty years ago. More recently, a study was published in

Change magazine about the relationship between faculty development and student learning. Rutz, Condon, Iverson, Manduca, and Willett (2012) stated that the underlying assumption of faculty development programs "is that when faculty learn more about teaching, they teach better, which in turn improves student learning" (p. 41). To test this assumption, the researchers tracked the effects of faculty development on student learning at two institutions. One institution was a moderately selective public university (Washington State University; WSU) and the other a small selective school (Carleton College). At both institutions, student learning was investigated by measuring institution-wide initiatives. Specifically, researchers studied critical thinking at WSU and student writing at Carleton College.

Frequency of attendance at faculty development programs served as an independent variable; faculty were categorized as being low users (attending an average of 2.2 events), high users (attending 1-3 additional events) or very high users (attending more than 3 additional events). The researchers found that, on average at WSU, the more events a faculty member attended the higher that faculty members' students' critical thinking scores were. However, these results were not statistically significant. Likewise, the researchers did not find any improvement trends in student writing ability at Carleton College.

The logic of this study parallels the current study. However, the faculty development experiences were not rooted in prior assessment data. Also, the researchers were not interested in the effect of a particular type of faculty development experience (e.g., course design institutes), but rather, the cumulative effect of faculty attendance at any faculty development program on student learning. Also, the study did not find

statistically significant results. Focusing on an intensive faculty development experience, as opposed to uncoordinated, voluntary, faculty development programs, may yield more compelling results than the study presented by Rutz et al. (2012). For example, if the faculty development opportunities focused on how to integrate critical thinking in the class – as opposed to any generic faculty development exercise – there should be a greater impact on students' critical thinking skills.

Course Design Institutes. Within faculty development, teaching programs typically fall into two categories: those that focus on specific pedagogies (e.g., team-based learning) and those that facilitate course design. Fink's (2003) popular integrated model walks faculty members through designing courses by first analyzing situational factors (e.g., class size, time of day, student characteristics). Next, he encourages faculty members to create learning objectives for their course, considering the goals they hope students achieve beyond the course. Following objective development, faculty members are encouraged to create formative and summative assessments that align directly with these objectives. Finally, learning activities are developed that bridge the objective and assessments and help students achieve the learning outcomes. This process is also called "backward course design" and is described by other faculty development authors as well (e.g., Hansen, 2011). Interestingly, the backward design process is similar to programmatic assessment (i.e., begin by establishing outcomes, create assessments, then learning activities).

The key to the Fink (2003) model is alignment. That is, objectives, assessments, and activities within a course must be aligned to create an optimal learning experience for students. In higher education, many faculty become experts in a particular domain of

content; however, they do not necessarily learn how to create and teach aligned, integrated courses. Hansen (2011) explained the traditional course design model: to begin, faculty members first create or adapt course content; next, they plan their assignments and tests; following, they determine their grading procedures; and lastly, they create course objectives (Hansen, 2011). In this framework, planning is focused on the instructor and not the students, who are passive learners (Blumberg, 2009).

Course design experiences help faculty align their courses and also adopt learner-centered approaches to teaching (Blumberg, 2009). Learner-centered teaching "…emphasizes a variety of different method types that shift the role of instructors from givers of information to facilitators of student learning or creators of an environment for learning" (p.3). Within her chapter, Blumberg (2009) built a body of evidence supporting the benefits of learner-centered teaching including its impact on student learning and motivation.

Level Problem

Generally, course design efforts help faculty members to design or redesign their individual course sections. Most courses are a part of an academic program (Hansen, 2011). Within that program, there could be multiple faculty members teaching the same course; and, the course *sections* often vary drastically. The syllabus frequently provides the only insight into learning that occurs within such course sections. Syllabi from multiple sections associated with one course could vary in learning outcomes, assessment methods, textbooks, and even content. Unfortunately, academic programs typically lack course blueprints causing problems with curricular cohesion (Hansen, 2011).

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.

Therefore, communication across sections and courses would be beneficial (Hansen, 2011). 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.

The idea of faculty developers assisting in curricular efforts is not new. Indeed, in the 1980s, faculty development offices were called to help with general education and academic program reforms (Gaff & Simpson, 1994). Currently, there are issues facing higher education that faculty developers can help address, such as "assessment of student learning and curricular innovations" (Ouellett, 2010, p.11). Faculty developers could help by facilitating discussions, providing evidence of the impact of previous curricular change, and assisting in the review of existing programs. Faculty developers can also help by facilitating backward design institutes that focus on learner-centered practices.

Obstacles to True Loop Closure

Given the demand for improved learning by accrediting bodies and the short supply of its evidence, it is important to review the factors that may impede the process of closing the loop. One such obstacle is fear. In particular, some faculty members fear that programmatic assessment will be used in personnel decisions, a practice now common in K-12 (Pepin, 2014).

In addition, faculty members lack the time, space, or incentive to systematically improve their programs. Faculty members are busy and undertaking a programmatic task of this magnitude is not in most academics' purview. Finally, many faculty lack the expertise to make systematic program changes since they are not necessarily trained to teach, though as previously discussed, they are experts in a particular content domain (Bok, 2013). Once faculty members feel competent as professors, they are reluctant to change or admit that their practice is not as sound as it could be (Pepin, 2014).

"Learning Intervention"

Course design institutes are offered at many institutions and take many forms (e.g., Cornell University, n.d.; Indiana University South Bend, n.d.; Stanford University, n.d.; Suffolk University, n.d.; Tufts University, n.d.). Though literature on the nature of these programs is absent, it appears that all institutes span several days and are intense in nature. Likewise, course design institutes tend to be voluntary and are designed for faculty members to design or redesign their individual courses.

For the present study, a small group of faculty went through a course design institute (i.e., jmUDESIGN) focusing on similar learning outcomes (ethical reasoning). During this institute, faculty members worked to infuse learning outcomes from a campus

program, The Madison Collaborative: Ethical Reasoning in Action, into their courses. At the same time, the faculty members learned about backward course design (Fink, 2003). Rather than focusing on an individually selected aspect of their course, faculty participants dedicated their course design time to infusing programmatic ethical reasoning learning outcomes into their courses.

jmUDESIGN. At our institution, the faculty development center facilitates an annual weeklong intensive course design institute called "jmUDESIGN" (James Madison University, n.d.). This institute is fashioned after Fink's (2003) backward design model, wherein faculty members develop significant learning outcomes that map to course assessments and activities. About 20 faculty members participate each year. Faculty members are divided into teams and have a group learning facilitator (GLF). Over the course of five days, faculty learn to create meaningful student learning outcomes, to design formative and summative assessments to measure those activities, and finally, to create learning activities that aid students in achieving the outcomes. Throughout the institute, there is a great emphasis on the alignment between outcomes, assessments, and activities in the course. The process is very similar to program assessment, but is focused on the course level.

On the first day of the institute, faculty members are encouraged to identify situational factors and create and articulate a "five-year dream." Situational factors refer to any contextual variable that a faculty member must consider when designing a course. For example, teaching via an online medium, teaching for the first time, teaching to students afraid of the subject, and teaching to a large classroom are all situational factors. Faculty must acknowledge these factors as they move forward. The five-year dream, on

the other hand, answers the question of what each faculty member would like students to know, think, or do *five years after the course*. Often, faculty will focus on the detailed content within a particular course; the five-year dream exercise helps members to focus on the most impactful learning that will occur throughout the semester.

On the second day of the institute, after defining their five year dreams, faculty generate course learning objectives that are specific and measurable. The following day, participants create formative and summative assessments aligned with these objectives. On the fourth day, faculty are exposed to several evidence-based pedagogies and are encouraged to adopt learning experiences that align with their objectives and assessments. On the final day, faculty members articulate their pedagogical choices made during the institute, share their work with others, and reflect on the experience.

The jmUDESIGN institute follows a detailed curriculum created by JMU faculty developers. The institute is facilitated using a parallel process and backward design principles were used to develop the institute. Thus, jmUDESIGN has objectives, learning assessments, and both formative and summative assessments. Participants experience a tightly aligned "course" as they learn about how to design their own.

While the main learning intervention is concentrated within a week, additional learning and work carries forward. During the institute, faculty focus on designing a module or unit for an upcoming course, which tends to translate into faculty members creating learning objectives, assessments, and activities for a one- or two-week period of their course. However, many faculty members opt to redesign their whole course, which takes additional time outside of the intensive week. Faculty developers are available to support these efforts.

Faculty participants experience jmUDESIGN in a small group of 4-5 individuals. Typically, teams are multi-disciplinary and comprise of members who focus on very different student learning outcomes. However, the faculty participants in this dissertation all focused on *the same learning outcomes* (i.e., ethical reasoning outcomes articulated by the Madison Collaborative program). Faculty members participating in course design with a common focus is a new idea.

Literature Review Summary

In higher education, many programs assess student learning. However, the majority of programs do not use the results from their assessment efforts to improve student learning. This is largely due to a lack of faculty time and expertise. Faculty development centers provide support in a variety of ways for individual course sections. For example, if a faculty member were interested in learning about a new pedagogical technique, developers would support him or her via consultations, workshops, and the like.

In order to improve student learning at the program level, faculty need focused time and instruction. This experience should equip program faculty with the time and guidance necessary to make a systematic program change that would enhance student learning. In this dissertation, five faculty members participated in jmUDESIGN focusing on a small, common set of learning outcomes on ethical reasoning skill development.

The Madison Collaborative

The author's institution is a member of SACSCOC, which requires a Quality Enhancement Plan (QEP). Specifically, SACSCOC states that the QEP exists for "…engaging the wider academic community and addressing one or more issues that

contribute to institutional improvement the plan should …describe a carefully designed and focused course of action that addresses a well-defined topic… related to enhancing student learning" (SACSCOC, n.d.).

At James Madison University, the QEP is titled, "The Madison Collaborative: Ethical Reasoning in Action" (James Madison University, 2013). The mission of the Madison Collaborative is to, "Prepare enlightened citizens who apply ethical reasoning in their personal, professional, and civic lives" (James Madison University, 2013, p. 22). At the heart of The Madison Collaborative are the Eight Key Questions (8KQ), which provide a framework for students to use when faced with an ethical dilemma.

Seven learning outcomes aligned with the 8KQ drive The Madison Collaborative's programming; these outcomes are listed below in Table 1. Five of the seven outcomes are cognitive and two are attitudinal. The Madison Collaborative is a systematic university-wide program, complete with an intervention plan. Specifically, planned interventions include:

1) A 75-minute session during freshmen orientation entitled, "It's Complicated: Ethical Reasoning in Action."
2) An online interactive experience spanning eight weeks
3) Programming in residence halls
4) Curricular interventions for faculty including coverage of the Eight Key Question framework (James Madison University, 2013).

The first intervention occurred for the first time during the summer of 2013, while the second intervention is being piloted with a small group of students during the 2014-2015 academic year. Intervention three started occurring in 2013 and affects both

freshmen and sophomores. The intent for the fourth intervention is for faculty to infuse ethical reasoning into their courses; this intervention consists of a program that spans one day of a weeklong university-wide professional development event hosted by the faculty development center.

While the Madison Collaborative is not a traditional academic program (e.g., Biology, B.A.), it is a program guided by student learning outcomes. Also, there is assessment evidence that suggests students are not currently proficient at ethical reasoning. Thus, there is a need to improve this area.

Ethical Reasoning – the Eight Key Questions

The Madison Collaborative is based on a variety of traditional philosophical and psychological approaches to ethical reasoning. Although philosophers and theorists tend to apply one philosophical approach to ethical reasoning, the Madison Collaborative uses casuistry, a method that integrates approaches (William Hawk, personal communication, June 20, 2013). The Eight Key Questions (8KQ) encourage students to weigh eight considerations of ethical reasoning based on six philosophical perspectives. The purpose of this endeavor is to help students to make better decisions by engaging in the ethical reasoning process. The 8KQ are described below.

Outcomes. The outcomes question asks students to consider, "What are the short-term and long-term outcomes of possible actions?" (James Madison University, 2013, p.19). This utilitarian approach has roots in John Stuart Mill's notion to promote the greatest good for the greatest number of people (Smith, 2014). In addition to weighing outcomes as Mill described (i.e., greatest good for greatest number), students also are encouraged to evaluate both short-term and long-term outcomes prior to making

an ethical decision. This perspective can be interpreted in other ways as well (e.g., in terms of karma as an outcome).

Fairness. The fairness question asks students to consider, "How can I act equitably and balance all interests?" (James Madison University, 2013, p.19). This perspective is based on John Rawls' *A Theory of Justice* (1971). The idea of fairness is that everyone should be treated equally and is especially applicable to situations where societal inequalities arise (Smith, 2014).

Authority. The authority question asks students to consider, "What do legitimate authorities (e.g., experts, law, my god[s]) expect of me?" (James Madison University, 2013, p.19). The consideration of authority figures' expectations when reasoning through an ethical dilemma is related to Piaget's *Theory of Cognitive Development* (1932) and Kohlberg's *Theory of Moral Development* (1969). In the Madison Collaborative, students are encouraged to identify what constitutes a legitimate authority when considering this perspective.

Rights. The rights question asks students to consider, "What rights (e.g., innate, legal, social) apply?" (James Madison University, 2013, p.19). Students are encouraged to consider if any rights of any person are compromised during an ethical dilemma. The rights perspective is grounded in Kant's (1797) principles on duties of rights and virtues.

Liberty. The liberty question asks students to consider, "What principles of freedom and personal autonomy apply?" (James Madison University, 2013, p.19). Students are encouraged to consider if an ethical situation impedes on their own or others' personal freedom or autonomy. The liberty perspective is also grounded in

Kant's (1797) principles. Liberty is a right. Because liberty is a pronounced right in the United States, it stands alone in the 8KQ framework.

Responsibilities. The responsibilities question asks students to consider, "What duties and obligations apply?" (James Madison University, 2013, p. 9). As with liberty and rights, the responsibilities consideration comes from Kant's (1797) philosophy. This perspective is often considered when the person in an ethical situation feels a responsibility to decide a certain way, or they experience a duty to others based on his or her role or perhaps profession.

Empathy. The empathy question asks students to consider, "How would I respond if I cared deeply about those involved?" (James Madison University, 2013, p.19). The empathy perspective is based on Gilligan's (1982) work. Gilligan (1982) considered the perspective of women in moral reasoning and challenged Kohlberg's justice-centric theory.

Character. The character question asks students to consider, "What actions will help me become my ideal self?" (James Madison University, 2013, p.19). This final question is grounded in Aristotle's *Nicomachean Ethics* (Smith, 2014). In this work, Aristotle defines character as something that sets one individual apart from another. One's character is a result of the development of one's virtues. This perspective encourages students to envision their ideal self and how they might respond to the ethical situation from that perspective.

8KQ Example. To reiterate, students are asked to learn the 8KQ, then to weigh different relevant questions when faced with an ethical decision. For example, a student may see a hungry child steal food from a vendor. The student must then decide whether

or not she will turn the child in to the police. In reasoning through this decision, the student may weigh all eight key questions or just the perspectives she deems most relevant. In this example, the student chooses to consider outcomes, authority, empathy, and character.

Regarding outcomes, the student may predict the outcome of turning the child into the police (i.e., the child would be punished) or not (i.e., the child would not be punished and would be fed). Within this perspective, she may also consider long-term outcomes such as the impact of being identified as a criminal at an early age or the impact of stealing with no repercussions.

In this scenario, the student is also considering the authority perspective. That is, in our country, the law (a legitimate authority, in this case) states that it is illegal to steal. If the student is religious, she might consider a religious text that advises against stealing. Also in this example, the student considers empathy and character. The student may feel empathy for the hungry child, which may influence her decision. Likewise, the student might consider what her ideal self would do in this situation. After reasoning through multiple perspectives, the student would then make a decision that is aligned with their reasoning. The intent of the 8KQ is for students to have a framework to guide their decision making instead of relying on snap judgments or only considering one perspective (e.g., authority).

Assessment of Ethical Reasoning

There were two instruments used to assess students' ethical reasoning skills, an Ethical Reasoning Identification Test (ERIT) and an Ethical Reasoning Rubric. These assessments are described below.

ERIT. The first assessment, the Ethical Reasoning Identification Test (ERIT), was designed to assess the first two Madison Collaborative learning outcomes (i.e., identification and selection of the 8KQ). On the ERIT, students are given a series of ethical situations and are asked to select the key question that best aligns with a given scenario. The ERIT has strong psychometric properties. There was support for the unidimensional structure of ERIT scores (Smith, 2014). Moreover, omega (Green & Yang, 2009) estimates of latent reliability were 0.79 (Smith, 2014). There is also validity evidence for ERIT score inferences. ERIT scores are related to — but distinct from — SAT Critical Reading scores. Moreover, students who have had light exposure to ethical reasoning interventions perform slightly better than those who have had no exposure to ethical reasoning interventions (Smith, 2014).

Ethical Reasoning Rubric. The other instrument is the Ethical Reasoning Rubric, which is designed to measure the fifth and most complex outcome. Students are given an essay prompt where they are instructed to 1) identify an ethical situation or dilemma that they have personally experienced, 2) apply (weighing, and if necessary, balancing) the considerations raised by the 8KQ, and 3) provide a decision after reasoning through the dilemma. The Ethical Reasoning Rubric (see Appendix A) is applied to student essays. This rubric has five evaluation criteria: 1) ability to identify an ethical situation, 2) appropriate reference to key questions, 3) determination of key question applicability, 4) analysis of individual key questions, and 5) weighing the relevant factors and coming to a decision. The rubric was intentionally aligned with the assessment instructions and also the fifth Madison Collaborative learning outcome (i.e., students will evaluate courses of action by applying [weighing and, if necessary,

balancing the considerations raised by Key Questions] to their own personal, professional, and civic ethical cases). This rubric has five scale points: 0) insufficient, 1) marginal, 2) good, 3) excellent, and 4) extraordinary and includes behavioral anchors. Raters using the rubric may assign half points.

Faculty members familiar with the 8KQ framework are recruited to evaluate the ethical reasoning essays. At a minimum, faculty raters must have attended a workshop on the 8KQ. Faculty members who volunteer to rate essays are paid an honorarium to attend a rating session that occurs over one day and includes training and calibration (approximately 1.5 hours); more time is spent rating ethical reasoning essays.

Two raters (i.e., a rater team) evaluate each essay so that inter-rater reliability may be calculated using generalizability theory (G-theory). G-theory allows researchers to analyze sources of error variance (e.g., error due to raters or elements of the rubric). Preliminary generalizability studies (g-studies) conducted after the first essay rating session from 2014 indicated that most unique variance associated with essay scores was attributed to essay differences, which is desirable. More detail about the reliability of scores will be reported in the Method and Results section. Each rater team evaluated about 20 essays; essays were de-identified to maintain students' confidentiality.

Within this study, the ERIT and the Ethical Reasoning Rubric were used to measure students' ethical reasoning skills. An emphasis was placed on the ethical reasoning essay scores, which measured a more complex application of ethical reasoning.

Research Questions

To date, the literature does not include methodologically sound quantitative or qualitative studies evidencing improved learning at the program level. The existing

examples tend to be anecdotal and short on details. Specifically, they do not provide a rich description of the faculty experience of going through an intense development learning intervention or of implementing new course components designed during said learning intervention. A new process is needed; one that encourages the use of results and provides the time, space, and guidance to do so. This process must integrate and infuse faculty development knowledge into program assessment frameworks. My proposed faculty development learning intervention was intended to meet this need. To this end, the following questions guided this study:

1. What is the effect of a program-level faculty development learning intervention on student's ethical reasoning identification ability? (quantitative-primary question)
    a. Hypothesis: Students enrolled in professors' courses who experienced the faculty development learning intervention will have higher ethical reasoning scores on the ERIT than a control group.
2. What is the effect of a program-level faculty development learning intervention on student's ethical reasoning skills? (quantitative-primary question)
    a. Hypothesis: Students enrolled in professors' courses who experienced the faculty development learning intervention will have higher ethical reasoning scores on the Ethical Reasoning Essay than a control group.
3. What is the experience like for faculty who participate in the development learning intervention? (qualitative- secondary question)

4. What is the experience of teaching newly designed ethical reasoning components like for faculty members? (qualitative- secondary question).
5. How could the faculty development learning intervention be improved? (qualitative- secondary question)
6. What results emerge from comparing the faculty experience qualitative data with the student outcome quantitative data? (mixed method- secondary question)

CHAPTER 3

Method and Results

Philosophical Foundation

As customary in mixed methods research, I begin with a discussion of my philosophical worldviews— beliefs and assumptions about knowledge that inform a mixed methods study (Creswell & Plano Clark, 2007). Creswell and Plano Clark (2007) posited four overarching worldviews: postpositivism, constructivism, participatory, and pragmatism. The postpositivist worldview tends to guide quantitative research, focusing on cause-and-effect relationships and testing pre-existing theories. The constructivist worldview guides qualitative research, building and constructing knowledge from data (e.g., interviews and observations). Participatory research advocates for participants to collaborate with them through the research process. Finally, pragmatism is focused on "what works" in applied settings and employs multiple methodologies (Creswell & Plano Clark, 2011, p. 41).

Mixed methods research combines quantitative and qualitative methods of inquiry. These two approaches to inquiry differ and are thought by some to be incompatible. Howe (1988) summarized these critiques and coined the term "incompatibility thesis": the belief that quantitative and qualitative methods should not be combined (Howe, 1988, p.10). Howe (1988) and Onwuegbuzie and Leech (2005) argued against the incompatibility thesis by highlighting similarities in the two methodologies. For example, both approaches use observations to address research questions. Likewise, researchers in both frameworks build safeguards into the research process to minimize confirmation bias. Because of such similarities, these authors argue that quantitative and

qualitative methodologies can and should be joined. Additionally, the two approaches can complement one another when combined (Feilzer, 2010).

Creswell and Plano Clark (2007) argued that it is acceptable and encouraged for mixed method researchers to have multiple worldviews. Thus, it is not unusual in a mixed methods study to see a researcher shift in worldviews at different stages. For the present study, I adopted an overarching worldview of pragmatism, which embraces a pluralism of methods (Onwuegbuzie, Johnson, & Collins, 2009). This choice is reflected in my decision to conduct a mixed methods study. I was interested in understanding how faculty experience jmUDESIGN and determining how to better the program in addition to measuring its impact.

While my primary stance was pragmatic, I shifted worldviews during different phases of the study. Specifically, I approached the quantitative data analysis as a postpositivist, the qualitative analysis as a constructivist, and during the study's integration phase I applied my pragmatic worldview again. I intentionally shifted paradigms at different stages to be true to each method's philosophical underpinnings. For example, I used an array of quantitative analyses (e.g., generalizability theory, calculation of Cronbach's alpha), all of which were created with the assumption that there is a truth in the population that can be inferred (i.e., a postpositivist assumption). Likewise, when analyzing the qualitative data I focused on describing the experience, not generalizing; a practice aligned with constructivism.

In sum, my overall research inquiry was pragmatic. I was interested in creating a new structure for higher education to facilitate learning improvement; in order to do so, I needed to know *what works* and *how.* Thus, my overall intent is to apply the results of

this inquiry, not to create new theories or generalize to a greater 'truth.' However, I wanted to stay true to each methodology's philosophical underpinnings and so I switched worldviews within each strand of inquiry (i.e., quantitative/postpositivism, qualitative/constructivism). For study integration, I reconciled the two strands using my pragmatic perspective.

Design

A mixed methods design was used, with an emphasis on the quantitative strand (i.e., student learning). The qualitative strand was also of interest because this was the first time faculty participated in jmUDESIGN focusing on a common set of learning outcomes. Specifically, I used an embedded mixed methods design; one methodology (i.e., qualitative) was embedded within another (i.e., quantitative; Creswell & Plano Clark, 2011)[1]. I refer to the quantitative strand as the "student impact study," which addresses research questions one and two. Research questions three, four, and five, were answered through qualitative inquiry; this strand is referred to as the "experience study." The embedded design was chosen because, in addition to learning about the program impact (on students), the qualitative component allows the researcher to understand the intervention (i.e., jmUDESIGN) experience for faculty.

Although useful, the embedded mixed method design is not without criticism. Plano Clark et al. (2013) summarized the critiques against embedded mixed method designs. Historically, mixed methodologists have expressed concerns that the qualitative component is often not robust or well conceptualized from the beginning of the study. Plano Clark and colleagues (2013) suggested considering the following connection points

[1] Creswell (2014) now calls this design an "intervention design."

to ease critics' concerns: development of the research questions; the data collection design; and data analysis, results, and interpretation.

In this dissertation, I considered both quantitative and qualitative strands when developing the research questions. I also considered when each strand of data was collected. A timeline of data collection that highlights both strands is depicted in Figure 2. The white boxes indicate qualitative data collection points (i.e., observation, journaling, and interviews) and the black boxes indicate quantitative data collection (i.e., rating of student essays).

Special attention was given to the *timing*, *weighting*, and *mixing* of the two approaches. Regarding *timing*, the main phases of data collection were during and after jmUDESIGN. As represented in Figure 2, qualitative data focusing on faculty were collected during and after the experience and quantitative student data were gathered during Week 13 of the fall 2014 semester.

Weighting refers to how much emphasis is placed on each strand of inquiry. In the current study, the most weight was given to the quantitative research question because the literature lacks evidence of learning improvement. Lastly, the data were strategically *mixed* at the interpretation phase (i.e., in answering research question six). Table 2 displays a summary of research questions and data sources.

Organization of Methods and Results

There were many data sources in this dissertation (i.e., ERIT scores, Essays Ratings, Observations, Interviews) and each data source was collected using a different methodology. Thus, for clarity, the methodology and results for each data source are presented in tandem and are organized by research question. The quantitative strand

methodology and results are presented first followed by the methodology and results of the qualitative strand. Study integration and a discussion of the study limitations follow.

Quantitative Strand - Student Impact Study

The student impact study was the emphasis of this dissertation addressing the two primary research questions:

1) What is the effect of a program-level faculty development learning intervention on students' ethical reasoning identification ability (as measured by the ERIT)?

2) What is the effect of a program-level faculty development learning intervention on students' ethical reasoning skills (as measured by the Ethical Reasoning Essays)?

Student Data Collection. Five faculty members attended the program-level faculty development learning intervention, jmUDESIGN. During this experience, these faculty members infused ethical reasoning into their courses. Due to a programmatic scheduling error, one of the jmUDESIGN participants did not have the opportunity to teach their re-designed course, although the other four did. Thus, student data were collected from only four of the five participants.

To answer research questions one and two, I invited students from the four faculty members' courses to take an ethical reasoning assessment. I hypothesized that these students would perform better on the ERIT and the Ethical Reasoning Essays than a control group because they were enrolled in courses that intentionally taught the 8 Key Questions (8KQ) framework (James Madison University, 2013). Students recruited from the four faculty members' courses were in the *treatment* group. Students who took ethical reasoning assessments during Assessment Day are referred to as the *control* group. Data

collection details for each group follow. All students were enrolled at James Madison University in Virginia.

Control Group. James Madison University holds two Assessment Days per year. On these days, classes are canceled and the university collects student learning data for accreditation purposes. All students are required to attend Assessment Day, both as entering freshmen and again when they are mid-way through their sophomore year. On Assessment Day, students engage in 2.5 hours of testing on student learning measures used by General Education, Student Affairs, and university wide initiatives (e.g., the Madison Collaborative). Students are randomly assigned to rooms with differing test configurations based on the last two digits of their student identification cards. If students do not participate in Assessment Day, a hold is placed on their account and they cannot register for classes.

Trained proctors monitored students as they worked on their assigned assessments to ensure they were on task. Each assessment was allotted a specific amount of time (e.g., 60 minutes) and students were not allowed to progress to the next assessment until the duration of time ended.

The random sample of students, who were assigned to take ethical reasoning assessments on Assessment Day, served as the control group. Freshmen in this sample had exposure to the 75-minute ethical reasoning session, *It's Complicated*, during the university's orientation (which occurred the day before they took the test). Sophomores in this sample did not experience *It's Complicated*, although it is possible that they had exposure to Madison Collaborative programming through other means (e.g., through residence hall programming).

Treatment Group. As previously mentioned, students in the treatment group were enrolled in one of the four jmUDESIGN faculty participants' classes. Data collection for this group was held outside of the classroom, in the evenings of Week 13 of the fall semester. Three of the four faculty participants offered extra credit to students for attending a data collection session. For these three faculty members (i.e., Professors 1-3), students were invited to participate one week in advance of the sessions.

To incentivize participation, pizza was provided after the one-hour data collection sessions. When students arrived at a session, they were given a consent form, which asked if they consented to participate in the research study and if they consented to release their SAT scores (to be used as a covariate). Students could earn participation points as long as they remained for the entire session, regardless of whether or not they consented to the research. If a student did not consent, their data were discarded and not used. Students' extra credit points were solely based on participation and not performance. This intentional configuration was designed to mirror the low-stakes environment of Assessment Day (i.e., students are only required to participate).

Data were collected in tow rooms: a computer lab and a traditional classroom. Students in the computer lab responded to the ethical reasoning essay prompt and students in the classroom completed the Ethical Reasoning Identification Test (ERIT). More students were randomly assigned to the ethical reasoning essay prompt than to the ERIT because it measures a higher cognitive ability. Recall the Ethical Reasoning Rubric measures *reasoning* abilities and the ERIT measures *identification and selection* of the 8KQ.

The fourth professor, who taught a small class, allowed me to administer the ERIT and Ethical Reasoning Essays during one hour of his/her class. The professor did not provide additional credit to students. Regardless of data collection setting, the test administration procedure mirrored that of Assessment Day. Table 3 displays student participation information by Professor. In total, 192 students participated in this study. The majority of participants were enrolled in courses taught by Professors 1 and 3 (85%). Likewise, more student data were gathered on the Ethical Reasoning Essays than the ERIT, by design. Finally, 88% of students who attended the testing session consented to participate in this study.

Propensity Score Matching. To answer research questions one and two, I needed to make meaningful comparisons between the treatment and control group average scores. An obstacle to such comparisons can be underlying differences in the groups due to factors other than the intervention. There is reason to believe that such pre-existing differences exist because students in the treatment group were not randomly assigned to courses—as students in the Assessment Day sample were randomly assigned to take ethical reasoning assessments. That is, students in the treatment group may have opted to take a certain professor's course for a variety of reasons (e.g., reputation of the professor). In research design, random assignment into both the control and treatment group is desirable to balance out student differences on a variety of characteristics and reduce selection bias (Shadish, Cook, & Campbell, 2002).

Selection bias occurs when one group systematically differs from another group on confounding variables. It is good practice to attempt to balance sources of selection bias (Yanovitzky, Zanutto, & Hornik, 2005). One way to reduce such bias is to use

propensity score methodology (Stuart & Rubin, 2007). This approach allows researchers to adjust for selection bias effects in a more robust way than traditional methods (e.g., multiple regression; Yanovitzky et al., 2005). Propensity score methodology matches individuals from a control group to a treatment group by balancing identified confounding variables.

Stuart and Rubin (2007) defined a best practice in propensity score matching: researchers should first choose covariates, then select a distance measure, select matches, diagnose matches, and finally analyze the data. Thus, I organized my propensity score matching process in congruence with these guidelines. I used MatchIt (Ho, Imai, King, & Stuart, 2006) in R software version 3.1.2 to estimate all propensity scores (R Core Team, 2014).

***Student Opinion Survey*.** Mostly demographic variables were used as covariates in this study. Additionally, students' test-taking motivation scores, as measured by the Student Opinion Survey (SOS; Sundre, 2007), were used as covariates because data were collected in low-stake settings. That is, students had incentives to participate, but not to perform well. Students in low-stakes testing environments tend not to do their best (Wise & DeMars, 2005). Thus, the SOS, which can be found in Appendix B, was administered after the ethical reasoning assessment to measure student test-taking motivation. The SOS measures students' test-taking motivation via two subscales: Effort and Importance (Sundre, 2007). The five effort items were written to measure the self-reported amount of effort students dedicated to the test at hand and the five importance items were written to measure students perceived importance of doing well on the tests. Both effort and importance scores were also used as covariates.

RQ1- ERIT. To answer research question one, "What is the effect of a program-level faculty development learning intervention on students' ethical reasoning identification ability (as measured by the ERIT)?" a comparison was made between the control and treatment group after matching said groups using propensity scores.

***Participants*.** In Fall 2013, 504 students took the ERIT on Assessment Day. Likewise, 794 students took the ERIT on Spring 2014 Assessment Day. Thus, the control group contained 1,293 students. The treatment group contained 69 student responses.

***Data Characteristics*.** The ERIT consists of 50 items; scores range from 0-50, with 50 indicating a perfect score. Students responded to the SOS using a 5-point Likert scale where 1-Strongly Disagree and 5-Strongly Agree. Each subscale of the SOS consisted of five items, and thus the range for these measures was 5-25, with higher scores reflecting higher values on the construct (i.e., effort or importance). Of note, two items were reverse-scored prior to creating a total-score. Finally, there is support for a two-factor structure that is invariant across males/females and across computer-based and paper-pencil testing modalities (Thelk, Sundre, Horst, & Finney, 2009).

Cronbach's alpha (α) is a measure of reliability, or internal consistency, which ranges from 0 to 1, with higher values indicating higher reliability (Meyer, 2010). Cronbach's alpha was calculated for the ERIT, SOS-Effort, and SOS-Importance. Reliability values for all data sources associated with the ERIT are presented in Table 4, along with descriptive statistics.

***Procedure*.** As Stuart and Rubin (2007) recommended, I began by selecting covariates that may be related to treatment assignment (i.e., students selecting a particular

class). I had access to the following variables: race, gender, SAT Math scores, SAT Critical Reading scores, students' test-taking motivation (effort and importance), and whether or not students experienced *It's Complicated.* Of note, MatchIt (Ho, Imai, King, & Stuart, 2006) cannot handle missing data, therefore all missing data were deleted.

Select Covariates. Recall that for the treatment group, students were asked to consent to the release of their SAT scores. For the ERIT, about half of the students *did not* release this information. Thus, if SAT scores were used as a covariate, the sample size for the treatment group would have reduced from 69 to 33. Because I did not want to lose half the sample, I decided against including SAT scores. Thus, for the ERIT, the following covariates were used: race, gender, students' test-taking motivation (effort and importance), and whether or not students experienced *It's Complicated.*

Select a Distance Measure. After identifying the covariates, I selected a distance measure, which evaluates the extent to which cases are similar in covariate values (Stuart & Rubin, 2007). I used a logistic regression model to estimate propensity scores (i.e., distances). Using logistic regression, each case was assigned a propensity score that indicated the probability of that case being assigned to the treatment group, given the set of defined covariates (Yanovitzky, Zanutto, & Hornik, 2005).

Select Matches. Once propensity scores were estimated, I used the one-to-one greedy nearest neighbor matching algorithm with a 0.20 caliper to select matches. One-to-one nearest neighbor matching selects a match for each treatment case from the control group based on the proximity of their propensity scores (Austin, 2011a). If several control group cases have equally close propensity scores the match is selected at random. The addition of a caliper specifies that the match must be within a specified distance of

the treatment case. I took Austin's (2011b) recommendation of setting the caliper to 0.20, selecting cases within 0.20 standard deviations of the propensity score. During the matching process, two treatment cases were lost due to missing data and all but one case was matched to the control group resulting in 66 students in the treatment group and 66 students in the control group.

Diagnose the Match. After the matched samples were generated, I determined the quality of the match. Austin (2011a) recommended evaluating standardized mean differences on covariates and also evaluating the balance between the treatment and control group across the entire dataset. Table 5 includes descriptive statistics and the standardized mean differences for the matched samples. Standardized mean differences less than 0.10 have been regarded as negligible differences (Normand, et al., 2001). A few covariates (i.e., Race, Gender, and SOS Importance) had standardized mean differences greater than 0.10, though only slightly. Density graphs for the two continuous covariates, effort and importance, are displayed in Figure 3.

To understand the overall balance, I first created a ratio of the treatment group propensity score variance compared to control group propensity score variance. Ideally, this value will yield a value near 1.0, indicating similar variances. For this match the ratio was 0.95. Recall that this ratio should be near 1.0, so this value suggests balance. Likewise, I also created graphs to depict the overall match. Visual inspection of the jitter plot in Figure 4 and the histograms in Figure 5 suggest balance in the matched groups.

Results. An independent samples t-test was conducted to determine if the treatment group (M = 31.72, SD = 8.34) differed from the control group (M = 33.58, SD = 8.28) on total ERIT scores. The two groups were not statistically significantly different

from one another, $t(128) = 1.28$, $p = .21$. Likewise, the effect was in an unexpected direction with the control group scoring higher on the ERIT than the treatment group ($d = -0.22$).

***Interpretation*.** I hypothesized that the treatment group would score higher on the ERIT than the control group (i.e., students randomly assigned to take the ERIT on Assessment Day). However, the results suggest no statistical difference between the two groups.

RQ2- Essays. Like with research question one, propensity score matching was conducted prior to comparing the treatment and control groups in order to answer research question two, "What is the effect of a program-level faculty development learning intervention on students' ethical reasoning skills (as measured by the Ethical Reasoning Essay)?"

***Participants*.** In Fall 2013, 133 students wrote an ethical reasoning essay during Assessment Day that was evaluated using the Ethical Reasoning Rubric. Likewise, 42 students wrote Ethical Reasoning Essays during Spring 2014 Assessment Day. Thus, the control group consisted of 175 student essays and the treatment group consisted of 122 student essays.

***Data Characteristics*.** Students were given the ethical reasoning essay prompt and were instructed to write a minimum of 250 words during 60 minutes. Faculty members rated the essays during one of two essay rating sessions (i.e., one session for the control and one session for the treatment group). Descriptive statistics on the average scores by element are reported in Table 6. Although element ratings are reported here for descriptive purposes, the overall average score was used to represent ethical reasoning

skills. The control group average score was close to *marginal* on the Ethical Reasoning Rubric (see Appendix A), and the treatment group average was between *marginal* and *good.*

Rater Teams. Two raters are randomly assigned to a team. Five teams rated the control group essays during Summer 2014 and four rater teams evaluated the treatment group essays in January 2015. During the latter rating session, five teams were supposed to evaluate essays but one rater did not attend as planned and thus the other raters evaluated the 20 essays assigned to the fifth team after completing their assigned essay. Thus, reliability information is only presented for the four rater teams.

The overall ratings assigned by Team are graphed in Figure 6. These average ratings can only be compared for relative purposes since each rater team scored different essays. To evaluate rater harshness, a common, or "anchor" essay, was rated by all teams in 2014 and 2015. Ratings for the anchor essay are depicted in Figure 7. The 2014 raters gave this anchor essay an average of 3.3 (SD = 0.30) and the 2015 raters assigned an average of 3.4 (SD= 0.40). Thus, the 2015 raters may have been more lenient.

Generalizability Theory. For the essay ratings, generalizability theory (g-theory) was used to parcel out sources of error (e.g., due to rater harshness). Conceptually, g-theory is a statistical method that estimates the dependability of behavioral measurements (e.g., essay ratings; Shavelson & Webb, 1991). That is, how well does a person's score generalize to a universe of other possible scores they might have received under certain conditions?

First, I defined the g-theory design. G-studies include facets, which are analogous to factors (independent variables) in traditional analyses (e.g., t-test, ANOVA). The

present study had four facets: person (i.e., the essay), rater, team, and items (i.e., the criteria used on the rubric). The design of the overall g-study is ((pxr):t x i). Or, persons' essays are crossed with the rater who evaluated the essay; this rater is nested within a rater team. Finally, these facets are crossed with "items"; thus, the same rubric criteria were used by all raters and were applied to all essays.

The team facet could not be studied because only one team evaluated each essay. This is not problematic because raters were randomly assigned to teams and raters did not speak to their rater-teammate about the essays. Thus, on average, any effects due to teams should be 0. The overall design was not fully crossed,[2] that is, every rater did not evaluate every essay (although every rater did use all rubric elements). Within teams, however, the design is fully crossed because both raters evaluated the same subset of essays using all rubric criteria. Thus, to determine the reliability of essay scores I conducted separate D-studies[3] for each team using GENOVA (Brennan, 2001). All facets were treated as random and not fixed facets.

In G-Theory, the G coefficient is akin to Cronbach's alpha in classical test theory (Meyer, 2010). In this case, the G coefficient is an estimate of how consistently two raters rank order student essays. As with alpha higher values reflecting higher reliability. Specifically, Hoyt (2010) offers that 0.80 and higher reflect "good dependability of scores" (p. 152) and 0.70-0.79 reflect "marginal dependability" (p. 152), although Hoyt (2010) notes that these are just rules of thumb and estimates must be interpreted in the

[2] Fully crossed designs exist when each facet occurs with every other facet and the object of measurement (in this case essays).
[3] D-studies, or "Dependability Studies" estimate variance components and produce reliability coefficients.

context of the study. The G coefficient is used to determine reliability of *relative* decisions—that is, how well do students compare to one another? The G-coefficient is calculated using *relative error,* which includes all variance components that interact with the object of measurement (i.e., all variance components except for the item variance).

The Phi (Φ) coefficient is a reliability estimate associated with making *absolute decisions.* That is, how well do students perform relative to a standard? The Phi coefficient is calculated using *absolute error,* which includes all variance components except for the object of measurement (i.e., the item variance is considered error). Because more terms are used in the absolute error calculation than the relative error, the Phi coefficient will usually be lower than the G coefficient.

Variance components, G and Phi Coefficients, and standard errors for the control group are reported in Table 7 and the same information for the treatment group is reported in Table 8. Figure 8 compares the reliability coefficients for the two groups. The person variance component in Tables 7 and 8 reflects the variability due to differences in essay ethical reasoning quality; this variance is desirable. Other variance components should be low. The rater variance reflects rater harshness and the item variance indicates that some rubric elements were more difficult to score well on than others. The interactions reflect more complex sources of errors: some essays differed by item (i.e., person x item), some raters scored certain essays more harshly than others (i.e., person x rater), and some raters scored certain rubric elements differently than others (i.e., rater x items). The last variance component, "person x rater x item, error" (read person by rater by item confounded with error) is the last source of variance that contains all additional error variance that could not be parceled out.

For most rater teams, the person variance component contributed the most to the total variance (which is desirable). Most teams also have a sizable amount of variance due to person x rater interactions, meaning that raters evaluated different essays differently. Also of note, rater team 3 in 2015 had an unusually large proportion of variance (56%) due to the rubric criteria (i.e., some criteria were easier to obtain higher values than others).

The G and Phi coefficients are around the acceptable range of 0.70 (Hoyt, 2010). Thus reliability of ratings is acceptable, although there was variability among teams and some were low. In addition to sources of error contributing to low reliability estimates, the low variability among essay scores is also causing these estimates to be low (i.e., most scores were low). Further, the rubric criteria (i.e., items) were treated as random. An argument could be made that the criteria are fixed given that the rubric was designed specifically for the Madison Collaborative's ethical reasoning framework (i.e., the rubric elements might not be exchangeable). Thus, this facet was treated as random to be conservative. Had it been fixed, the variability associated with the items would be pulled into the person variance increasing score dependability.

Reliability. G-theory was used to estimate the dependability of ratings. Cronbach's alpha was estimated to evaluate the internal consistency of student motivation scores. For the control group, effort ($\alpha = 0.77$) and importance ($\alpha = 0.81$) had high reliability estimates. Likewise, effort ($\alpha = 0.85$) and importance ($\alpha = 0.82$) reliability estimates were high for the treatment group.

Procedure. As with the ERIT, I followed Stuart and Rubin's (2007) guidelines for propensity score matching: choose covariates, select distance measures, select matches,

diagnose matches, and analyze the data. Likewise, I had the same covariates to choose from: race, gender, SAT Math scores, SAT Critical Reading scores, students' test-taking motivation (effort and importance), and whether or not students experienced *It's Complicated.* Again, missing data were deleted because MatchIt cannot process incomplete datasets.

Select Covariates. Missing data for SAT scores in the treatment group was an issue. Like the ERIT, many students chose not to release their SAT scores. Including SAT scores would have reduced the treatment sample size from 122 to 69; thus to maintain statistical power, I decided against including SAT as a covariate. As with the ERIT, the final following set of covariates was used: race, gender, students' test-taking motivation (effort and importance), and whether or not students experienced *It's Complicated.*

Select a Distance Measure. The distance measure indicates how similar cases are on a set of covariates (Stuart & Rubin, 2007). As with the ERIT analysis, I used a logistic regression model to estimate propensity scores.

Select Matches. I used a one-to-one greedy nearest neighbor matching algorithm with a 0.20 caliper to select matches. The matches were created using MatchIt (Ho, Imai, King, & Stuart, 2006) in R software version 3.1.2 (R Core Team, 2014). This process resulted in 107 matched cases. Twelve treatment cases were left unmatched due to the inclusion of the 0.2 caliper.

Diagnose the Match. To determine the quality of the match, I created standardized mean differences on covariates, which are displayed in Table 9. Standardized mean

differences were all less than 0.10. Density plots for effort and importance (i.e. the only continuous covariates) are in Figure 9.

I evaluated overall balance by first creating a ratio of treatment group propensity score variance to compare to control group propensity score variance; this ratio was 1.10 which is close to 1.0 (i.e., perfect balance). Finally, I created a jitter plot (see Figure 10) and histograms illustrating the distribution of propensity scores across the two groups (see Figure 11) for the overall dataset. A visual analysis of these plots jitter plot and histograms reveals that the essay data were well balanced on all covariates used.

***Results*.** Using the matched sample generated from propensity score matching, I conducted an independent samples t-test to see if the treatment group (M = 1.49, SD = 0.63) differed from the control group (M = 1.07, SD = 0.40) on total ethical reasoning rubric scores. The difference was statistically significant, $t(212) = -4.70$, $p < .001$. The effect size (d = 0.80) was large. Recall that a 1 on the rubric represents *marginal* and a 2 represents *good.* Therefore, interpreting the raw effect size reveals that the average control group essay scores were *marginal* and the treatment group essay scores were between *marginal* and *good.*

***Interpretation*.** As hypothesized, students in the treatment group scored higher on the Ethical Reasoning Rubric than students in the control group. The standardized effect between these two groups was large and the raw effect size is meaningful.

Qualitative Strand –Experience Study

The experience study is the secondary focus of this dissertation and addressed three research questions:

3. What is the experience like for faculty who participate in the faculty development learning intervention?
4. What is the experience of teaching newly designed ethical reasoning components like for faculty members?
5. How could the faculty development learning intervention be improved?

Faculty Data Collection. To understand the faculty experience, I collected multiple forms of data. Specifically, I observed participants during the five-day jmUDESIGN experience, I asked participants to journal daily, and lastly, I invited each faculty member to be interviewed: once after jmUDESIGN and once after infusing ethical reasoning into their course.

***Faculty Participant Characteristics*.** Five faculty members participated in jmUDESIGN, with the intent of infusing ethical reasoning content into their courses. Each participant was personally invited to attend jmUDESIGN by a Madison Collaborative representative. Most of these faculty members were selected because they had previously attended a training session on the Madison Collaborative and the 8KQ; one faculty member did not receive prior training but was teaching a course on ethical reasoning. Four of the five jmUDESIGN participants were provided an honorarium by the Madison Collaborative for their participation; however, one participant was affiliated with the Madison Collaborative and did not receive the honorarium.

Two of the faculty participants were male and three female. One participant was from the College of Business, two were from the College of Education, and the remaining two participants were from the College of Arts and Letters. Only one of the five participants had previously attended the jmUDESIGN institute in the past.

Observations. Each day, I observed the five participants. I assumed the role of an observer-participant. That is, although I was there to observe the faculty participants, I shared their setting, at times even seating myself at their table. Unlike the participants, I did not redesign a course during the program. When conducting observations, Merriam (2009) suggested focusing on one of the following elements at a given time: the physical setting, the participants, activities and interactions, conversation, subtle factors, and your own behavior. Because jmUDESIGN had a set curriculum I paid less attention to the setting and activities (these were determined by the program) and more attention to the participants, their interactions, conversations, and body language.

During the jmUDESIGN curriculum, there are times when participants are passive (i.e., listening to a presentation) and there are also highly interactive times when participants are encouraged to engage with their group. I captured observational data during both passive and active times. I intently observed at different times of the day for 1-2 hours, although I attended most of the entire institute.

Daily Journals. In addition to observing participants during the day, I also asked the five participants to send me reflective journals at the end of each day. The journal prompt included the same question each day: What was the jmUDESIGN experience like for you today? I intentionally left this question open and focused on the experience. The jmUDESIGN curriculum included full 8-hour days and faculty members had homework assignments each night to complete. Thus, while all participants intended to journal, some were unable to or missed a few days due to the heavy workload and other life commitments. Specifically, only two participants provided a journal response every day,

another participant provided four out of five, the fourth participant wrote a journal response three of the five days, and the last participant did not journal at all.

Interviews. Observation and journaling occurred during jmUDESIGN. However, I was also interested in the sustaining or changing effect of the experience on faculty members. Thus, I scheduled two interviews with each faculty participant after jmUDESIGN ended. The first interview was held within a month of the conclusion of jmUDESIGN and the second interview was held after the fall semester ended. Because one faculty participant did not teach his/her course in the fall semester due to scheduling conflicts, only four of the participants were interviewed after the semester. All five participants consented to participate in this study.

Instrumentation. The summer interview questions (see Appendix C) mainly focused on the jmUDESIGN experience and the fall interview questions (see Appendix D) focused mostly on the implementation experience. Each interview question was viewed as a data source and all data sources were aligned with one of the three qualitative research questions. Table 10 shows the alignment of qualitative data sources to research questions.

Phenomenology. Phenomenology is a philosophical lens and a methodology. Edmund Husserl founded phenomenology in the early 20th century as a means to break away from philosophical abstractions and focus on actual lived experiences (Lichtman, 2012). Thus, a phenomenological study asks, "…what is this or that kind of experience like?" (van Manen, 1990, p. 9). The aim of such a study is to discover the "internal meaning structures" of lived experiences (van Manen, 1990, p.10). In other words, what is the *essence* of an experience? In phenomenology, one can never fully reduce the

human experience; rather, only apply a rich description of the experience. Unlike other methodologies (e.g., quantitative) the goal is *not* to generalize the experience of a few to that of many. The goal is to discover and describe a lived experience.

There are no prescribed methodologies to conducting phenomenological studies. van Manen (1990) states that while there is no method, there is tradition that guides the phenomenologist that stretches back in history. Thus, the body of phenomenological scholarship serves as a set of guides for practice.

***Bracketing*.** In phenomenology, bracketing—or deliberately suspending one's belief in order to study a phenomenon—is commonplace. However, Lichtman (2012) states that the idea of bracketing is overly simplified, as one can never truly remove bias. Although she does not believe one can fully set aside prior beliefs, she finds value in writing about one's experience with the phenomenon. Thus, as opposed to bracketing, in the true sense, I positioned myself prior to data analysis. The practice of positioning is common in qualitative studies.

***Positioning*.** I am familiar with jmUDESIGN and was a past participant. My jmUDESIGN experience was very positive. Further, I work at the university-wide faculty development center that developed jmUDESIGN. In this position, I assess faculty-learning outcomes, including the outcomes of the jmUDESIGN program. Thus, I know the outcomes and curriculum of the program very well.

My familiarity with this program had advantages and disadvantages. One advantage was that the program was familiar to me, so during observations I could focus on the participants rather than the setting. However, my familiarity with jmUDESIGN may also be a limitation as participants may hesitate to speak unfavorably about the

experience during the interview phase of this study. Such familiarity also plays into the intersubjective experience I had with participants.

***Intersubjectivity*.** Intersubjectivity theory suggests that experiences are mutually shaped (Stolorow & Atwood, 1996). For the current study, I was part of the experience with participants (i.e., I experienced the program and the interviews with them). During these experiences, I naturally developed reactions that feed into the intersubjective nature of studying an experience. I tried to maintain awareness of possible points of bias due to intersubjectivity during the analysis phase. In phenomenology, intersubjectivity is necessary for the researcher to understand the experience (van Manen, 1990). Unlike in quantitative research, bias is not only a limitation but also a fundamental component of the experience.

Analysis Stages. Using NVivo 10 for Mac (2014) I employed the following methodology for each research question. First, I open-coded the data for initial categories or themes. During this process, I kept the appropriate research question in mind (e.g., what is the jmUDESIGN experience like?). More weight was given to interviews; observations and journals were used to supplement the interview data source.

Following this initial coding process, I re-organized the codes based on content (e.g., joining similar codes). During this second phase, I also ensured that at least two of the participants' data were contributing to a code. I did not want any one persons' experience to overpower the connection of experience across participants. Next, I applied horizontalization; that is, I laid out all codes within a particular research question "flat" disregarding the frequency of references within a code. During this process, I evaluated each code individually determining its contribution to the research question. This process

yielded my final set of themes for each research question. Finally, I delved within each theme to make meaning of each facet of the experience.

Validity. In the quantitative paradigm, the term 'validity' has generally been agreed upon in the measurement framework (AERA, APA, NCME, 1999). However, there is debate within the qualitative paradigm over the term. In 1985, Lincoln and Guba presented the term *trustworthiness* to describe the reliability and validity of qualitative data. From this perspective, trustworthiness included data credibility, transferability, dependability, and confirmability. More recently, Morse, Barrett, Mayan, Olson, and Spiers (2008) criticized Lincoln and Guba's perspectives on validity stating it focused too much on the outcome of the analysis rather than the process. Citing Kvale's (1989) definition of validity as a process "… to investigate, to check, to question, and to theorize" Morse et al. argue that the term validity has a place in qualitative research (p. 19).

Morse et al. (2008) recommended shifting responsibility from external reviews (outcome based) to the researchers. Specifically, the authors recommended researchers employ verification strategies. In the current study, I strived for *methodological coherence;* the first verification strategy described by Morse et al. (2008). Methodological coherence ensures congruence between the research question and the method. I ensured coherence by 1) aligning interview questions to overarching research questions (see Table 10) and 2) constantly focusing on the *experience.*

In addition to focusing on the validity of the process, I also adhered to Creswell's (2013) recommendations. Specifically, I clarified my bias as a researcher in the positioning section, and I included negative cases during the analysis. Finally, because

this is a dissertation, I logged my steps throughout the process, which are susceptible to external audits by committee members.

RQ3- jmUDESIGN Experience. Research question three is, "What is the experience like for faculty who participate in the faculty development learning intervention?" The following interview questions, all asked during the summer interview after jmUDESIGN, pertained to this research question (as outlined in Table 10):

- What products did you create?
- What was it like learning about course design while simultaneously trying to infuse Ethical Reasoning into your course structure?
- Describe the best parts of the jmUDESIGN experience?
- What was the jmUDESIGN experience like for you?

***Analysis Details*.** As outlined in the data analysis section, I began by open-coding the summer interviews, observations, and journal entries. This initial coding resulted in 45 codes. Following the initial codes, I reorganized the data to include 33 codes. Of these, 13 codes had adequate representation from at least two participants to be considered during horizontalization. During the horizontalization process, I considered all 13 codes, regardless of weight (i.e., number of codes or data sources tied to the code) and focused on the codes that most reflected the research question. Five aspects of the experience were identified that best represent the jmUDESIGN experience. Figure 12 depicts these themes. All themes reflect a triangulation of the three data sources.

***Results*.** I discovered five themes of the jmUDESIGN experience: it is a focusing experience, the group aspect is important, the program is a learning experience, overall it is positive, and it is overwhelming. Each aspect is described below.

Focusing Experience. Three participants mentioned how the jmUDESIGN experience was a focusing experience; specifically, there were three interview references and one journal reflection reference. Faculty participants indicated that they appreciated having the time and space to focus. Professor 2 said, "…it was… a valuable focusing activity for me. Focusing on what I consider to be more significant important things in terms of my teaching." Likewise, Professor 3 stated, "It's one of those things that you think, 'it'll be really good for me to redo things in my class or introduce this approach in my curriculum.' But this [jmUDESIGN] really makes you stop and do it… otherwise time can just get away because other things are screaming more loudly."

Group Experience. This was the first jmUDESIGN group to focus on a common set of learning outcomes (i.e., ethical reasoning). During the summer interview and in their journals, four of the five participants mentioned the group aspect as one of the best parts of the jmUDESIGN experience. Interview data and journal responses revealed two facets to the experience: they enjoyed the group and they appreciated the opportunity to brainstorm since they were focused on common goals. Professor 1 said, "I did really enjoy our group. I think that because we were all working in the general [same] area, there was a fair amount of…. Camaraderie." Professor 3 stated that the common focus gave the group a special dynamic, which "…provided opportunities to brainstorm and inspire one another." I also observed many positive interactions among participants; they often helped each other, joked, and worked together. I did observe one philosophical disagreement about grades that came up between two participants a few times, but generally all interactions were positive and the group appeared to get closer over the five days.

Learning Experience. All five participants mentioned at least one thing they learned from the jmUDESIGN experience during their interviews and in their journal responses (if they submitted them). Responses focused on specific content and skills taught during jmUDESIGN (e.g., the importance of course alignment, the distinction between summative and formative assessment). Professors 2 and 4 spoke about the importance of focusing on their "five year dream;" as one said, "The constant theme that kept this in front of us and kept reminding me of the real end… was the reference to the five year dream. How does the learning objective align with the five year dream?" A more general quotation captures the overall nature of learning from the experience, "I feel like I've learned more practical approaches in the past three days than I have in as many years." These quotes were from participant interviews.

Positive Experience. All five participants, during interviews and through journals, said the overall experience of jmUDESIGN was positive. One participant said, "…positive….I never felt like, 'Oh my God I have to go there tomorrow.' I looked forward to going." Professor 3 stated, "The entire week was excellent and I feel like I have been able to come up with some good methods to incorporate the 8KQ into my course." Generally, participants discussed enjoying the institute because it was valuable, they were productive, the program was well designed, they enjoyed learning, and they liked the social aspects of the experience. During my observations, I saw participants frequently laughing and joking with one another. For the most part they were attentive and seemed relaxed.

Overwhelming. All participants also described the jmUDESIGN experience as overwhelming. Through interviews, journals, and direct interaction with me during the

institute, four of the five participants described at least one life event that co-occurred with the institute making it difficult for them to complete their homework and give the institute their full attention. As such, the participants reported feeling overwhelmed. Specifically, one participant said, "The days are long and I'm tired when I get home!" Another said, "… life did happen for me and I didn't feel that there was any room to accommodate … so in the middle I had to say, 'Alright, you know I'm going to make my choices of my own.'" Participants found it especially difficult to fully engage with the institute, especially in the evenings, because of life factors that were competing for their time.

RQ4- Teaching Experience. Research question four is, "What is the experience of teaching newly designed ethical reasoning components like for faculty members?" The following interview questions pertained to this research question (as outlined in Table 10):

- What challenges do you foresee in implementing the segment of your course that you redesigned?
- Have you implemented the redesigned components of your course? If so, please describe the experience
- What was the experience of teaching ethical reasoning like for you?
- What were the challenges of teaching ethical reasoning?

***Analysis Details*.** Initial coding resulted in 24 codes. Following the initial codes, I reorganized the data to include 17 codes. Many of these codes were sparse, and thus all 17 were used for the horizontalization process, which resulted in two overarching aspects of the teaching experience.

***Results*.** The participants discussed teaching ethical reasoning in many ways. Nevertheless, I identified one common theme among three of the faculty members who taught ethical reasoning: that the experience was fun.

Fun. During the second interview, three participants described the implementation experience as "fun." Two participants used the word "fun" and another said, "I really love doing it. I love discussing it, setting the stage, and, um, it set into motion this whole, uh, process which was unexpected…" Thus, although I did not observe their teaching, participants described the experience in the classroom as being enjoyable.

RQ5- Improvement. Research question five is, "How could the faculty development learning intervention be improved?" The following interview questions pertained to this research question (as outlined in Table 10):

- How could your experience have been improved?
- Recommendations for a jmUDESIGN totally dedicated to ethical reasoning?
- Can you think of any support that would have made teaching ethical reasoning in your course easier

***Analysis Details*.** This pragmatic question was analyzed using the same coding procedure outlined previously. Initial coding of interviews resulted in 17 codes. All codes were unique and not re-organized. Many of these codes were reported by only one participant and thus were not considered during horizontalization.

***Results*.** Horizontalization resulted in three recommendations for jmUDESIGN improvement; all recommendations focused on the 8KQ and not course design (i.e., the current focus of jmUDESIGN). Specifically, participants wanted to learn from people

who had previously taught with the 8KQ framework, they wanted more 8KQ specific resources, and they wanted more time with the 8KQ content.

Learn From Those Who Have Taught 8KQ. Four of the five participants indicated that they wanted to learn more from faculty members who had previously infused the 8KQ into their courses. These faculty members knew a person who had taught a course infused with ethical reasoning and all suggested it would have been helpful to learn from this person. Professor 1 stated, "…first and foremost, it would have been really nice for [other professor] to… give us a bit more content related help."

More 8KQ Resources. Three participants recommended having resources about the 8KQ content infused into jmUDESIGN. Currently, the institute is focused on designing aligned courses, and not on the content of courses. However, these participants felt that such resources would be helpful for their particular situation given the common focus. Professor 3 noted he/she wanted "some resources for [our] particular table… For instance, the person who was leading our table wasn't even trying to incorporate ethics." Thus, this professor was also suggesting that it would be helpful to have a group lead facilitator who is familiar with the 8KQ framework.

More Time for 8KQ. Finally, two participants suggested that the institute build in dedicated time to discuss the 8KQ. Thus, the learning experience would emphasize course design *and* the student learning outcome content. Professor 1 stated in an interview that, "…it would have been a little bit nicer if there was a little more ethics." Such time could be used to fully grasp the content and then apply it to one's course.

Implementation Fidelity. Implementation fidelity is a process whereby one evaluates the degree to which a planned intervention matches actual implementation

(Gerstner & Finney, 2013). In the context of this study, the examination of implementation fidelity addressed two questions: 1) did the faculty member implement the material they created during jmUDESIGN? 2) how much emphasis was placed on ethical reasoning in each class? To answer the first question, I asked the four participants what they created at jmUDESIGN and whether or not they implemented it. All four participants told me they implemented the materials they created during jmUDESIGN, indicating fidelity.

Regarding the second question, however, there was variability on the degree to which professors emphasized the 8KQ. Table 3 listed the four professors with the number of students enrolled in each of their classes. What follows is a description of ethical reasoning infusion in each course; this information came from Interview 2.

Professor 1 did not have formal training in the 8 Key Question (8KQ) framework prior to attending jmUDESIGN. That being said, this professor did teach a course called Ethical Reasoning; thus, the course was about ethical reasoning and students learned a variety of philosophical perspectives (e.g., Mills, Kant), but they did not put the perspectives together as the 8KQ framework does. During jmUDESIGN this professor made changes to his/her syllabus and created more formative assessments; this professor implemented these products as planned.

Professor 2 was intimately familiar with the 8KQ framework and also taught a course on ethical reasoning. This professor's course was intentionally aligned with the 8KQ. During jmUDESIGN, this professor created additional formative assessments, which were implemented.

Professor 3 developed ethical reasoning case studies and spent one class period discussing the 8KQ. Students also had virtual interactions with the 8KQ outside of class. An introduction to the 8KQ framework was early in the semester and the professor referenced the framework throughout the remainder of the semester. In fact, this professor always included at least one exam question that asked students to apply the 8KQ.

Professor 4 spent time during jmUDESIGN making room for the new content. Specifically, this professor added ethical reasoning case studies to the class. During implementation, the professor found that students voluntarily brought their own cases to class. Thus, students were regularly engaged with the 8KQ framework (more than the professor expected).

Integration

The last research question, a mixed methods secondary question, was "What results emerge from comparing the faculty experience qualitative data with the student outcome quantitative data?" Integration only occurred at the level of interpretation. Many mixed method studies will integrate the quantitative and qualitative data at the analysis level, creating joint displays (i.e., a visual display of how the two data types converge). However, because students were the focus of the quantitative study and faculty were the focus of the qualitative study, the only way to integrate the data sources would be to report the quantitative results by faculty member. This practice would violate my Institutional Review Board (IRB) approval, which assured faculty members that all data would be analyzed in the aggregate.

The quantitative results show that the treatment group performed better on the Ethical Reasoning Rubric than the control group, although this is not true for the ERIT comparison. Recall that the ERIT is aligned with the lowest level Madison Collaborative outcome (i.e., students will be able to state all 8KQ from memory). The Ethical Reasoning Rubric is aligned with the highest cognitive outcome—applying the 8KQ framework to situations in one's own life. The quantitative results are counterintuitive because the Madison Collaborative outcomes were built with a sense of cognitive order, assuming that one must know the 8KQ before applying them.

The qualitative results shed light on this odd finding. The jmUDESIGN curriculum began by asking faculty to focus on their five-year dream for students (i.e., what do you want students to know, think, or be able to do five years after taking your course?). The emphasis is on the higher-level, more sustaining skills that students will master as opposed to content particulars. jmUDESIGN also encourages participants to consider non-cognitive outcomes (e.g., values, motivation).

All participants mentioned learning new skills during the institute and two participants perseverated on the five-year dream aspect of the institute. Perhaps the focus on sustained learning worked, but at the expense of time spent on the foundational content knowledge (i.e., the 8KQ).

Pragmatically, the qualitative results suggest that the jmUDESIGN experience was positive and that faculty members used the products they created during the experience. This finding provides reason to continue adapting this assessment-faculty development model. Likewise, specific recommendations for future institutes were

gathered from the qualitative results. Had that component not been embedded in the primary quantitative design, a major pragmatic component would have been lost.

Limitations

In business, a proof of concept provides evidence that an idea is feasible (Proof of Concept, n.d.). This dissertation is a proof of concept that evidences that assessment and faculty development can be connected to improve programmatic student learning. Prior assessment results revealed that students were marginal at ethical reasoning. Thus, a group of faculty worked through jmUDESIGN together to infuse ethical reasoning into their courses. Subsequently, their students performed better on the Ethical Reasoning Rubric than students randomly assigned to take the same tests. Thus, connecting assessment with faculty development experiences to improve student learning is a viable solution to the use of assessment results problem.

However, this study has several limitations. First, although the Madison Collaborative is a program guided by student learning outcomes, it is not an academic degree program (e.g., Psychology, B.A.), which is the focus of traditional assessment efforts. In an academic program, the faculty group would certainly have a different dynamic than the current study given the social history they would have. Likewise, the content of focus would likely *not* be voluntary as it was with the Madison Collaborative. For example, if a psychology program wanted to improve students' statistical skills, the faculty members teaching that course would attend the faculty development experience focusing on that outcome; they *must* teach that outcome because it is part of the curriculum.

A second limitation is that the data in this study were nested. That is, students (Level 1) were nested within their professor's class (Level 2). Thus, students within one class may be more similar to one another than they are to students in another class. In an ideal design, hierarchical linear modeling would be used to account for this nested data structure (Raudenbush & Bryk, 2002). Such an approach would allow one to examine student learning differences that are dependent on individual faculty members.

Unfortunately, this approach is unfeasible due to practical restraints. Specifically, the sample size at the highest level (i.e., professors) is lower than 10, the bare minimum for the highest-level unit (Snijders & Bosker, 1999). Snijders and Bosker (1999) show that with ten units, fixed effects will be unbiased, but the standard errors for both fixed effects and variance components will be too small. Snijders and Bosker (1999) recommend 30 level-2 units (in this case, professors). Although hierarchical linear modeling cannot be conducted, the intraclass correlation coefficient (ICC) was calculated to determine the proportion of total variance that is due to between group variance (i.e., how much variance is due to students being nested within a particular professor's class?). This indicates the extent of dependence in scores due to which professor a student had.

The ICC for ERIT total scores was 0.04 and the ICC for Ethical Reasoning Essay scores was 0.02. Thus, 4% of the total variance in ERIT scores and 2% of Essay scores were due to the professor a student had. These proportions are quite small. Nevertheless, this variability could inflate alpha levels (Arceneaux & Nickerson, 2009).

There were several additional limitations associated with the quantitative, qualitative, and integration phases of the study. Regarding the quantitative results, the rater teams that evaluated the intervention group's ethical reasoning essays were slightly

more lenient than the control group's rater teams. Also, for propensity score matching, I used available covariates with sufficient data. Unfortunately, some students did not consent to release their SAT scores, which was a desirable covariate to use for balancing in propensity score matching.

It was also puzzling to discover no statistical difference between the control and treatment groups, but to find such a difference between these groups for the Essays. The mixed method integration allowed me to produce a hypothesis about this finding: perhaps the emphasis on the five-year dream encouraged faculty to focus on the more robust skillset as opposed to the foundational content. Of course, there are additional alternative hypotheses. Although students in the treatment and control group were matched on whether or not they experienced *It's Complicated*, there was a time difference for each group that might have caused a recency-effect. That is, the control group freshmen took the ERIT the day after receiving *It's Complicated* and students in the treatment group received the training two months prior to taking the assessment. Thus, it is possible that the control group performed higher than the treatment group because the *It's Complicated* training was likely fresh in their minds.

It is also possible that students do not necessarily need to be able to identify the 8 Key Questions prior to using them to reason through an ethical dilemma. Perhaps students can apply the framework without full knowledge of each perspective within the 8 Key Question framework. For example, a person may be able to drive a car and obey all rules on the road; however, the same person may not be able to pass a multiple-choice driver's test. Perhaps students are learning to reason without fully understanding the foundational content.

Qualitatively, I could have delved deeper in my interviews. I had the richest data for the first qualitative research question (i.e., what was the jmUDESIGN experience like?) partly because I had multiple data sources (i.e., interviews, observations, and journals). However, for the second interview, which focused on the teaching experience, I only had the single source of information and could have probed more to gather richer data. Finally, if I were to replicate this study I would have asked faculty members *if and why* they perceived jmUDESIGN to be effective.

Due to IRB restrictions, the data could not be disaggregated by professor's class. Future studies should include IRBs that allow for such disaggregation and report results by class, keeping the professors' identity anonymous. Disaggregation would allow mixed method integration to occur at the analysis phase (i.e., link qualitative and quantitative results by professor) in addition to the interpretation phase. Such information would be especially useful in the current study because faculty members had varying levels of prior experience with the 8 Key Questions. Likewise, professors varied in their implementation of the 8KQ (e.g., one faculty member taught an entire class on the framework and another introduced it and built upon the framework throughout the semester). Data disaggregation may have shed light on what, in particular, was effective about the jmUDESIGN experience.

Ultimately, the results suggest jmUDESIGN is related to increased student learning. However, we cannot disentangle *why* jmUDESIGN made an impact. Notably, the qualitative study showed that faculty members learned about course design during the formal curriculum, brainstormed about ethical reasoning with the other participants at their table, and used the week as a focusing activity. Was it the cumulative experience

that impacted student learning or did one factor contribute more than others? This question is left unknown.

Future Studies. Again, although there are a number of limitations to this study, there is evidence supporting that integration of faculty development with assessment is a worthwhile endeavor. Thus, future studies might apply this same model to an academic degree program, particularly a large one so that more faculty members participate and HLM can be used to appropriately model the data. `

Future studies might also attempt to disentangle the effect of the faculty learning intervention. Do faculty members just need time and space to develop curricula? Does the course design experience help? What would happen if faculty members already knew each other and worked together? These questions are testable and would likely benefit from mixed method designs.

Finally, future studies would benefit from data disaggregation and deeper exploration of implementation fidelity issues. Data disaggregation would allow for a clearer insight into *what worked.* Likewise, more targeted interview questions could be used to explore *why* certain classes performed higher than others. However, such exploration would come at the expense of faculty members' comfort. Faculty members may feel uncomfortable having their particular class compared to others even if their identity is kept anonymous. Addressing the limitations of the current study and building stronger studies in the future could yield larger effect sizes.

CHAPTER 5

Discussion

Higher education has many reasons to improve: to rise in global rankings, to address criticism, and to meet regional accreditation standards. Although higher education can demonstrate, via assessment mechanisms, that students are learning, we cannot provide evidence of learning *improvement* (Blaich & Wise, 2011). Such evidence is limited (Banta, Jones, & Black, 2010).

To address this issue, Fulcher et al. (2014) provided a simple model for evidencing student learning improvement: assess, intervene, re-assess. The authors noted that assessment professionals are trained to measure student learning, but rarely are they trained to support programs in making changes. Luckily, faculty developers are experts in facilitation, teaching, learning, and curriculum design and can assist programs and assessment practitioners in these efforts. Thus, in the simple model, assessment professionals can handle the assessment component, but they need assistance from faculty developers to achieve the intervention piece.

One reason faculty developers and assessment practitioners do not collaborate is because of the "Level Problem." That is, assessment efforts occur at the program level and most faculty development interventions are designed for course instructors focusing on a particular course section. For the simple model to be successful, faculty development experiences must occur at the *program* level. This dissertation provides a proof of concept for the assessment/faculty development partnership focusing on student learning outcomes. The particular program of focus was the Madison Collaborative: Ethical Reasoning in Action.

The Madison Collaborative

The Madison Collaborative is the result of James Madison University's Quality Enhancement Plan. The program is guided by the outcomes listed in Table 1. Essentially, the goal is for students to first learn the 8 Key Questions (8KQ) framework, and then use this framework to reason through ethical dilemmas they face in their life. Baseline assessment results suggested that students were adequate at identifying and applying the 8KQ framework (as measured by the ERIT) but they were not facile using the 8KQ to reason through ethical dilemmas.

This dissertation – through syncing of assessment and faculty development - sought to improve students' ethical reasoning skills. Thus, following baseline assessment, the Madison Collaborative, assessment practitioners, and faculty developers formed a partnership to do something differently (i.e., intervene). Specifically, the Madison Collaborative paid faculty volunteers an honorarium to participate in jmUDESIGN to infuse ethical reasoning into their courses. The faculty participants' students were invited to take an ethical reasoning assessment (either the ERIT or the ethical reasoning essay). These students comprised the treatment group and this assessment occasion constituted re-assessment.

To determine if the change in Madison Collaborative programming (i.e., the inclusion of jmUDESIGN) was an improvement, the treatment group was compared to the baseline group. The results suggested that students in the treatment group were better at ethical *reasoning* than the control group. However, the treatment group was not superior at identifying the basic premises of the ethical reasoning framework. This

finding was puzzling given that students should theoretically have command over the 8KQ prior to using them to reason through ethical dilemmas.

The qualitative strand of this dissertation explored the jmUDESIGN experience. This was the first time faculty members at JMU (and perhaps elsewhere) participated in course design while focusing on a common learning outcome. Thus, it was critical to understand this experience. The results suggested that while intense, the experience was positive and worthwhile. Faculty remarked that they learned many new things about teaching, and two participants perseverated on the ability to focus on one's five-year dream. This emphasis may explain the puzzling quantitative results—perhaps faculty members focused on the highest cognitive student learning outcome of the Madison Collaborative, forgetting to emphasize the foundational piece.

The Madison Collaborative has evidence suggesting that faculty members who infuse ethical reasoning into their courses during jmUDESIGN increased students' ethical reasoning skills. Likewise, the experience for faculty participants was positive. This particular finding is key from a pragmatic perspective because faculty buy-in is critical to improvement efforts.

Broader Implications

It is rare to encounter a program that can evidence learning improvement (Blaich & Wise, 2011). By applying Fulcher et al.'s (2014) simple model, this dissertation evidences learning improvement for the Madison Collaborative. The implications for the assessment field are vast. Currently, exhaustive efforts of assessment personnel are often placed on increasing assessment quality. Although assessment quality is important, it will not positively affect student learning – the goal of assessment. Thus, if higher education

is to shift from an assessment emphasis to one on learning improvement, then partnerships with professionals who can aid with the intervention component are essential (i.e., faculty developers). Assessment professionals must emphasize learning improvement and recognize that assessment is a necessary *tool* to achieve it—but not the answer in totality.

Learning improvement is not only beneficial to the assessment field, but also the faculty development domain. There have been recent calls for faculty development to employ more rigorous assessment methods (Chism & Szabo, 1997; Hines, 2009; Kucsera & Svinicki, 2010). To date, most faculty development assessment efforts emphasize faculty outcomes (Steinhert et al., 2006). Yet, a core assumption of faculty development is that by honing faculty skills students will learn more (Rutz et al., 2012). Even so, few studies can show that faculty development affects student learning. This dissertation provides evidence suggesting that jmUDESIGN makes such an impact, although more research is needed to untangle why this is the case. Nevertheless, by partnering with assessment practitioners, faculty developers are poised to gather student learning evidence to help demonstrate their impact.

Likewise, the qualitative results described the faculty experience of participating in a course design institute. These data are among the first to explore the course design experience, especially in this learning improvement context. The qualitative data supplemented the quantitative findings and provided direction for improving future course design institutes for programmatic learning improvement. Specifically, faculty participants suggested more time and resources for the student learning outcome content

be embedded into the experience. The mixed methods approach provided a more holistic understanding of this proof of concept.

Methodological Implications

This study used an embedded mixed method design to learn about the faculty experience in the learning intervention (i.e., jmUDESIGN) *and* determine its impact on student learning. Likewise, I used propensity score matching to create a balanced control group. Both methodological choices allowed for a robust understanding of learning improvement.

Often in assessment, quantitative methods are employed to directly measure student learning. Although this emphasis should not be devalued or replaced, assessment practitioners could benefit from incorporating qualitative strands into their inquiry. The integration of methods is especially helpful when studying learning improvement, a very new approach in higher education, which has much yet to be discovered.

Propensity score matching allowed me to balance the control and treatment groups on a set of covariates. A question that could easily be asked of learning improvement researchers is, "How do you know the treatment group wasn't just different from the control group?" Propensity score matching allows the researcher to answer this question in light of the covariates. Likewise, propensity score matching is preferred to standard methods of statistical control (Yanovitzky et al., 2005). I recommend the use of both mixed method designs and propensity score matching in future improvement studies, which align well with the emerging improvement science paradigm.

Improvement Science

Improvement science is an emerging paradigm that emphasizes improvement as opposed to experimental theory testing (Lewis, 2015). Plan-do-study-act (PDSA) cycles comprise the foundation of this science (Langley et al., 2009). These rapid cycles are guided by three questions: "What are we trying to accomplish? How will we know that a change is an improvement? What change can we make that will result in improvement?" (Lewis, 2015, p. 55). These three questions are similar to Fulcher et al.'s (2014) simple model; the parallels between the two models are presented in Table 11. The biggest difference between the guiding questions is the order of questions (i.e., the second and third questions are flipped).

Although this dissertation was guided by Fulcher et al.'s (2014) model, it has a strong parallel with the improvement science paradigm, which Lewis (2015) recommended education researchers consider. Researchers interested in programmatic learning improvement may benefit from further exploration in this new paradigm as the field progresses.

Conclusion

Assessment is a prevalent practice in higher education (Kuh & Ikenberry, 2009). Unfortunately, even the most interesting assessment findings do not prompt programs to change (Blaich & Wise, 2011). Therefore, assessment alone does not lead to learning improvement—something that would greatly benefit higher education. This dissertation explored learning improvement and provided a proof of concept for bridging assessment and faculty development—two offices that are rarely connected. This lack of connection evidences a structural pitfall to learning improvement. Higher education institutions

should intentionally connect these offices to facilitate and support learning improvement. Institutions should also invest in these initiatives (Banta & Blaich, 2011).

Although more research is needed to refine the learning improvement process, the initial evidence in this study indicates learning improvement can be achieved. If learning improvement were to proliferate in higher education much like assessment has in the past 20 years, the results would be profound for many stakeholders: students could learn more, faculty may engage in an enriching experience, assessment practitioners would provide due service to their learning improvement promise, faculty developers could demonstrate their impact, and institutions could champion improvements to regional accreditors. Finally, if the quality of higher education increased, the United States could regain its top position among global competitors.

There has been a great focus on assessment of student learning for the past two decades. With very few examples of learning improvement resulting from assessment, it is time to focus our energy on evidencing learning improvement – not just assessment. Learning improvement is the goal of assessment, after all. It's time we truly fulfill that purpose.

Table 1

Madison Collaborative Student Learning Outcomes

Cognitive	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 Key Question most consistent with the decision and rationale. 3. Given a specific scenario, students will identify appropriate considerations for each of the Eight Key Questions. Alternate approach: Students will be able to provide the specific considerations raised or rationale implied when applying every Key Question to an ethical situation or dilemma. 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 Key Questions. 5. Students will apply SLO 4 to their own personal, professional, and civic ethical cases. NOTE: Implied within this SLO is the students' ability to identify an ethical situation, based on the belief that the process of ethical reasoning increases discriminatory capacities. This will be addressed via the assessment rubric.
Attitudinal	6. Students will report that they view ethical reasoning skills as important. 7. Students will report increased confidence in their ability to use the ethical reasoning process.

Table 2

Matrix of Research Questions and Data Sources

Data Sources	RQ1- QUAN: ERIT Scores	RQ2- QUAN: Ethical Reasoning Essay Ratings	RQ3- qual: jmUDESIGN Experience	RQ4- qual: Teaching Experience	RQ5- qual: Program Improvement	RQ6- mixed: Study Integration
Student ERIT data	X					X
Student Essay data		X				X
Observations			X			X
Participant Journaling			X			X
Facilitator Journaling			X			X
Interview 1			X		X	X
Interview 2				X	X	X

Table 3

Treatment Group Sample Sizes

	Number Enrolled	Number Participated	Number Consented	Number ERIT	Number Essays
Professor 1	75	65	57	14	43
Professor 2	40	11	10	3	7
Professor 3	159	122	107	43	64
Professor 4	19	19	18	9	9

Table 4

ERIT Data Characteristics

	N	Cronbach's α	Mean	SD	Min	Max
Treatment						
ERIT	69	0.89	31.67	8.77	3.00	46.00
SOS-Effort	67	0.87	19.58	4.23	6.00	25.00
SOS-Import	67	0.78	15.81	3.96	5.00	25.00
Control						
ERIT	1298	0.87	33.75	7.88	4.00	48.00
SOS-Effort	1271	0.83	19.22	3.73	5.00	25.00
SOS-Import	1279	0.83	13.94	4.46	5.00	25.00

Table 5

Covariate Descriptives by Unmatched and Matched ERIT Groups

	Treatment (Unmatched) N=69		Control (Unmatched) N=1255		Treatment (Matched) N=65		Control (Matched) N=65		Standardized Mean Differences
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
White	0.83	-	0.86	-	0.85	-	0.92	-	-0.18
Asian	0.07	-	0.06	-	0.05	-	0.03	-	0.10
Black	0.01	-	0.05	-	0.02	-	0.02	-	0.00
Hispanic	0.03	-	0.05	-	0.03	-	0.02	-	0.00
American Indian	0.00	-	0.01	-	0.00	-	0.00	-	-
Pacific Islander	0.01	-	0.01	-	0.00	-	0.00	-	-
Gender	0.72	-	0.61	-	0.71	-	0.62	-	0.19
It's Complicated	0.81	-	0.39	-	0.83	-	0.85	-	-0.03
SOS-Effort	19.58	4.28	19.22	3.72	19.58	4.27	19.22	3.23	0.08
SOS-Importance	15.81	3.95	13.92	4.46	15.74	3.99	16.17	4.06	0.15

Note. Proportions are displayed for dichotomous variables (i.e., 1= presence of that variable and 0=absence of that variable). Gender was coded 1=Female, 0=Male. The SOS scores are continuous and include a mean and a standard deviation.

Table 6

Average Essay Ratings by Rubric Element

	Treatment (N=122)		Control (N=175)	
Rubric Element	Mean	SD	Mean	SD
A. Ethical Situation	2.36	0.87	1.94	1.16
B. Key Question Reference	1.44	0.92	1.13	0.94
C. Key Question Applicability	1.20	0.85	0.82	0.78
D. Ethical Reasoning: Analyzing Individual 8 Key Questions	1.23	0.90	0.86	0.82
E. Ethical Reasoning: Weighing the Relevant Factors and Deciding	1.15	0.89	0.90	0.83
Overall Average	**1.47**	**0.74**	**1.13**	**0.79**

Note. The scale is 0= Insufficient, 1=Marginal, 2=Good, 3=Excellent, and 4=Extraordinary.

Table 7

Control Group Variance Components in Ratings by Team

	Team 1	Team 2	Team 3	Team 4	Team 5
Person (%)	0.32 (34%)	0.43 (42%)	0.36 (22%)	0.52 (39%)	0.51 (48%)
Rater (%)	0.10 (10%)	0.00 (0%)	0.56 (34%)	0.04 (3%)	0.02 (2%)
Items (%)	0.09 (9%)	0.07 (7%)	0.14 (9%)	0.31 (23%)	0.20 (19%)
Person x Items (%)	0.04 (4%)	0.05 (5%)	0.09 (5%)	0.10 (7%)	0.05 (4%)
Person x Rater (%)	0.18 (20%)	0.21 (21%)	0.38 (23%)	0.12 (9%)	0.17 (16%)
Rater x Items (%)	0.01 (1%)	0.14 (13%)	0.00 (0%)	0.06 (5%)	0.00 (0%)
Person x Items x Rater, Error (%)	0.20 (21%)	0.11 (11%)	0.11 (7%)	0.20 (15%)	0.12 (11%)
G-Coefficient (Relative Standard Error)	**0.73** (0.35)	**0.77** (0.36)	**0.62** (0.47)	**0.84** (0.31)	**0.83** (0.33)
Phi- Coefficient (Absolute Standard Error)	**0.63** (0.43)	**0.73** (0.39)	**0.40** (0.47)	**0.73** (0.32)	**0.77** (0.39)

Note. Percentages of variance explained for variance components presented in parentheses.

Table 8

Treatment Group Variance Components in Ratings by Team

	Team 1	Team 2	Team 3	Team 4
Person (%)	0.51 (46%)	0.36 (26%)	0.18 (17%)	0.60 (46%)
Rater (%)	0.06 (5%)	0.33 (24%)	0.02 (2%)	0.14 (11%)
Items (%)	0.04 (3%)	0.19 (14%)	0.61 (56%)	0.20 (16%)
Person x Items (%)	0.08 (8%)	0.05 (4%)	0.08 (7%)	0.13 (10%)
Person x Rater (%)	0.25 (23%)	0.14 (10%)	0.07 (6%)	0.08 (6%)
Rater x Items (%)	0.02 (2%)	0.11 (8%)	0.02 (2%)	0.02 (1%)
Person x Items x Rater, Error (%)	0.15 (14%)	0.22 (16%)	0.11 (10%)	0.14 (11%)
G-Coefficient	**0.76** (0.40)	**0.78** (0.32)	**0.75** (0.24)	**0.88** (0.28)
Phi Coefficient	**0.72** (0.44)	**0.53** (0.56)	**0.48** (0.44)	**0.76** (0.43)

Note. Percentages of variance explained for variance components presented in parentheses.

Table 9

Essay Covariate Descriptives by Unmatched and Matched Groups

	Treatment (Unmatched) N=122		Control (Unmatched) N=175		Treatment (Matched) N=107		Control (Matched) N=107		Standardized Mean Differences
			Mean	SD	Mean	SD	Mean	SD	
White	0.76	-	0.85	-	0.84	-	0.84	-	0.00
Asian	0.07	-	0.05	-	0.06	-	0.05	-	0.03
Black	0.04	-	0.06	-	0.05	-	0.05	-	0.00
Hispanic	0.02	-	0.06	-	0.03	-	0.04	-	-0.04
American Indian	0.00	-	0.02	-	0.00	-	0.00	-	-
Pacific Islander	0.02	-	0.02	-	0.02	-	0.02	-	0.00
Gender	0.61	-	0.57	-	0.62	-	0.64	-	-0.03
It's Complicated	0.90	-	0.76	-	0.90	-	0.89	-	0.02
SOS-Effort	20.41	3.07	19.66	2.91	20.37	3.77	20.19	2.90	0.07
SOS-Importance	16.18	4.09	15.64	3.96	16.32	4.23	16.05	4.11	0.07

Note. Proportions are displayed for dichotomous variables (i.e., 1= presence of that variable and 0=absence of that variable). Gender was coded 1=Female, 0=Male. The SOS scores are continuous and include a mean and a standard deviation.

Table 10

Matrix of Qualitative Questions and Data Sources Organized by Research Question

Qualitative Questions	Interviews (Summer)	Observations	Journals	Interviews (Fall)
Research Question 3 – jmUDESIGN Experience				
What products did you create?	X	X	X	
What was it like learning about course design while simultaneously trying to infuse Ethical Reasoning into your course structure?	X		X	
Describe the best parts of the jmUDESIGN experience?	X		X	
What was the jmUDESIGN experience like for you?	X	X	X	
Research Question 4 – Teaching Experience				
What challenges do you foresee in implementing the segment of your course that you redesigned?	X			
Have you implemented the redesigned components of your course? If so, please describe the experience				X
What was the experience of teaching ethical reasoning like for you?				X
What were the challenges of teaching ethical reasoning?				X
Research Question 5 – Improvement				X
How could your experience have been improved?	X			
Recommendations for a jmUDESIGN totally dedicated to ethical reasoning?	X			
Can you think of any support that would have made teaching ethical reasoning in your course easier?				X

Table 11

Comparison of Fulcher et al.'s (2014) model to Improvement Science

Fulcher et al.'s (2014) Learning Improvement Model	Langley et al.'s (2009) Improvement Science Paradigm
1. **Assess** Learning Outcome of Interest	1. What are we trying to accomplish?
2. **Intervene** at the Program Level	3. What change can we make that will result in improvement?
3.**Re-Assess** to Determine if Change is Improvement	2. How will we know that a change is an improvement?

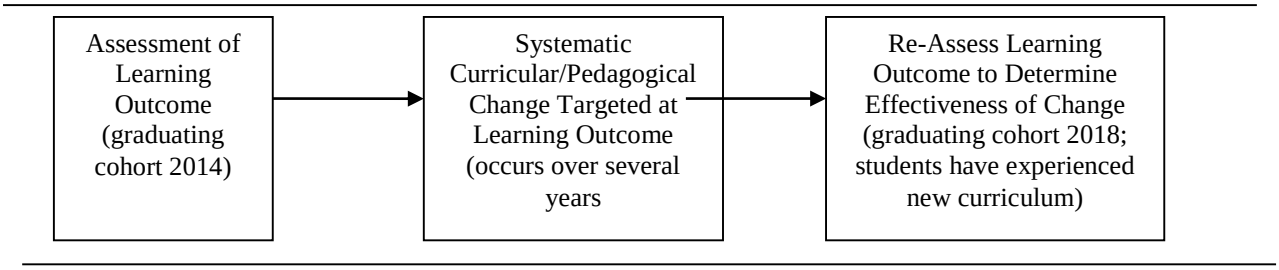

Figure 1. The Simple Model for Learning Improvement

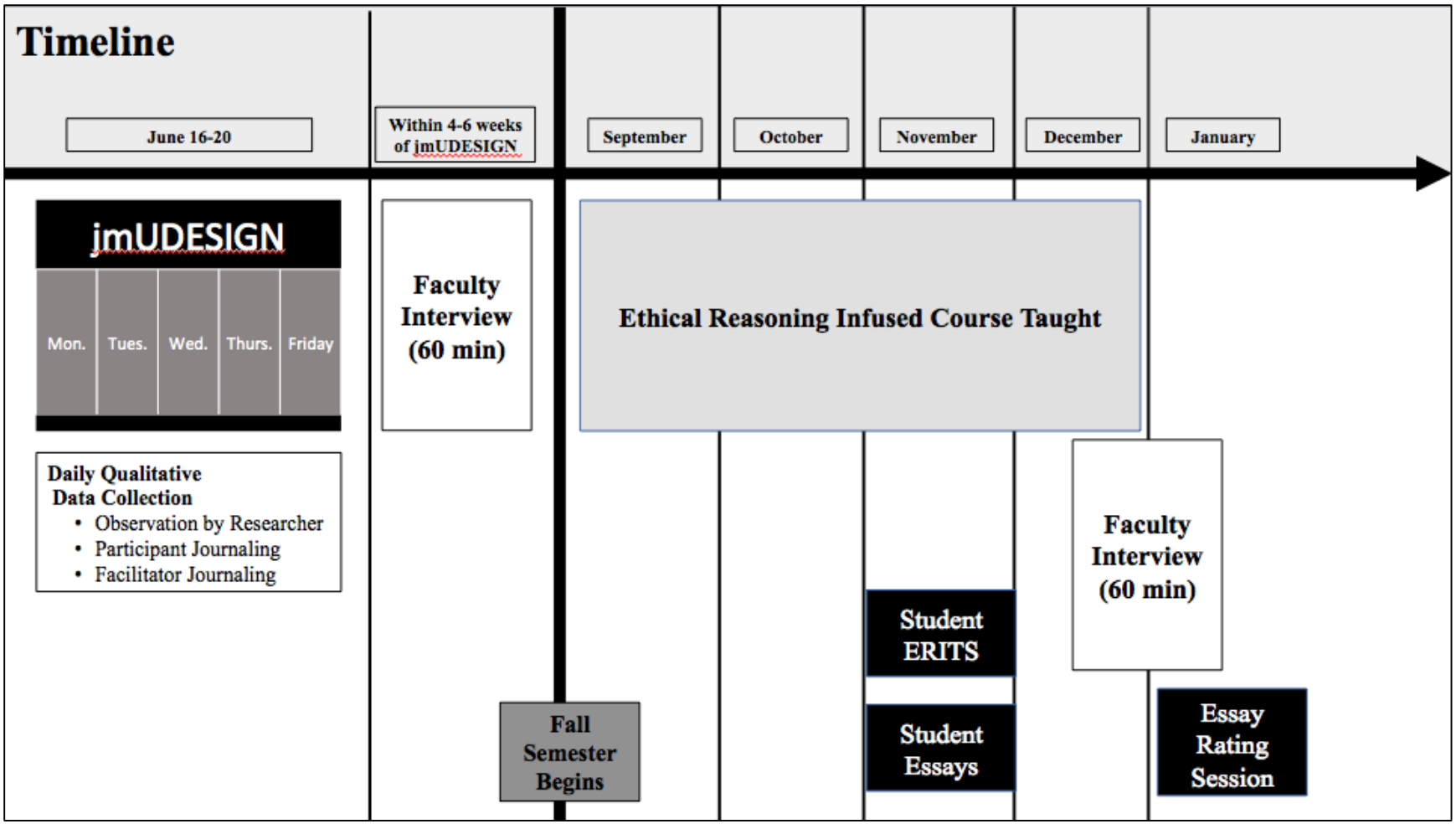

Figure 2. Data Collection Timeline

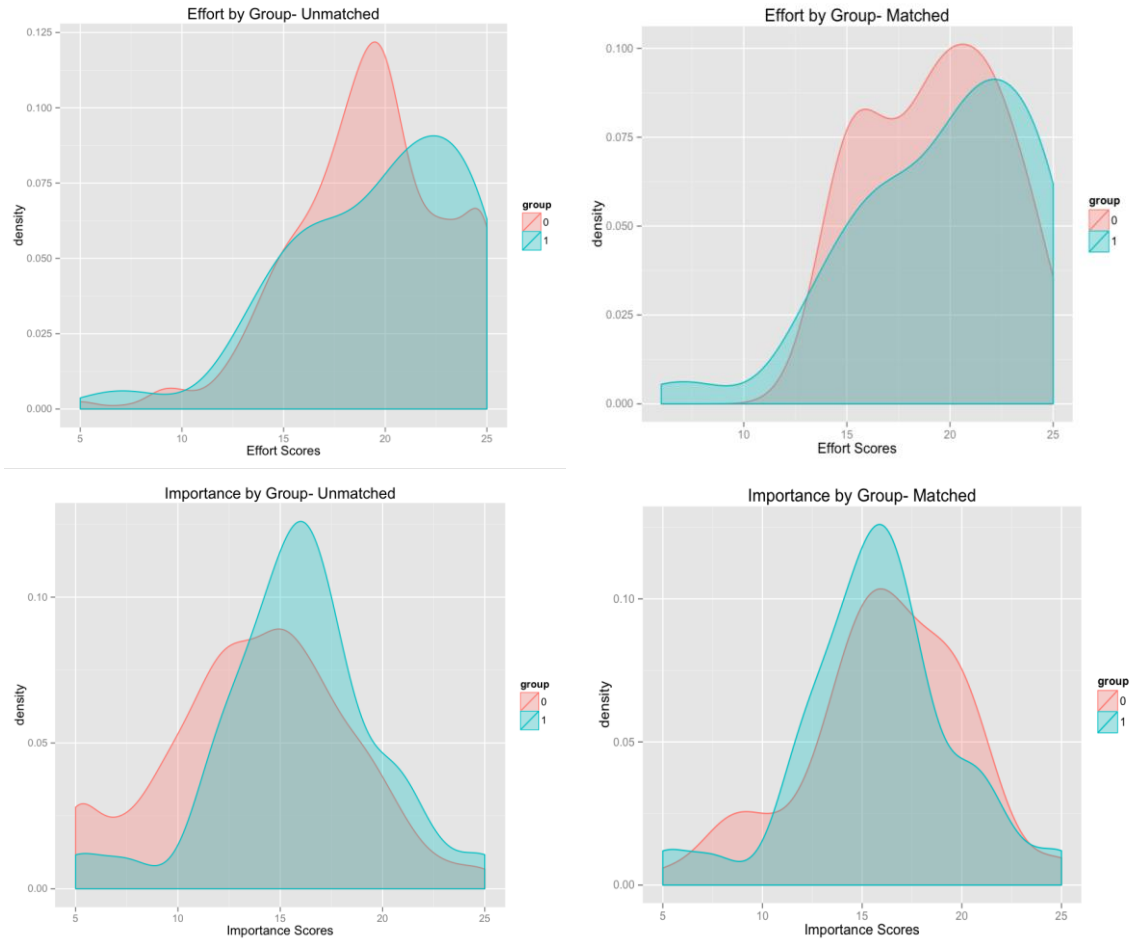

Figure 3. Effort and Importance for Treatment and Control Groups- ERIT

Note. N=65 for each group. The graphs display the cumulative density function for the Effort and Importance covariates after matching.

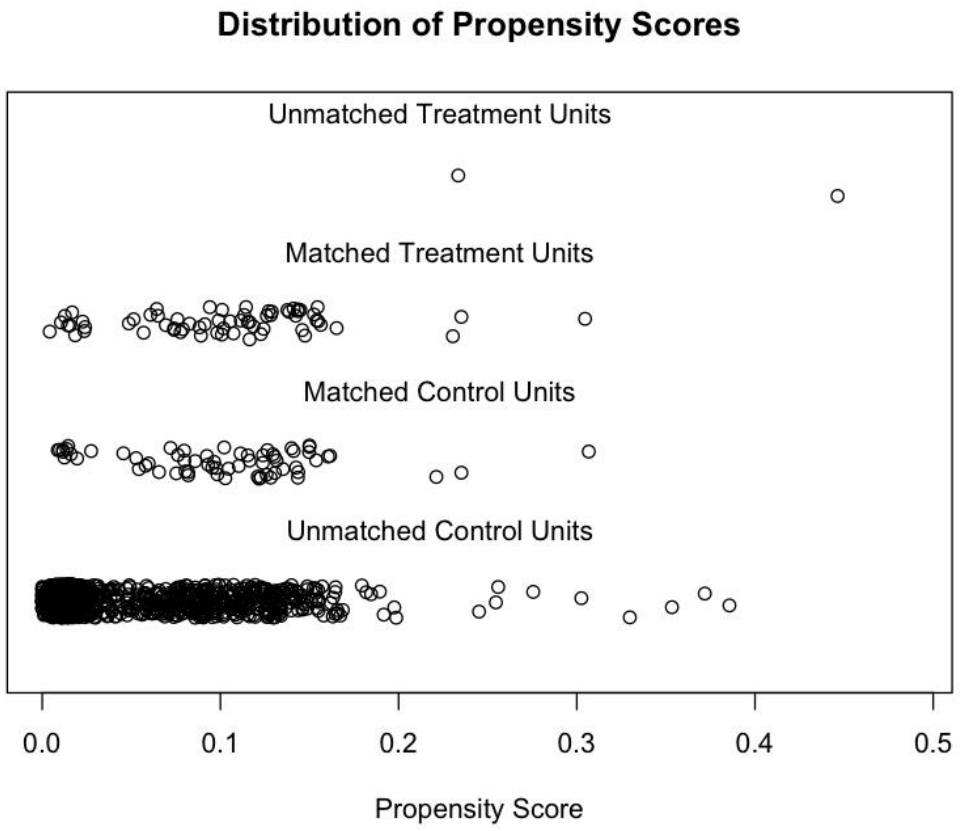

Figure 4. ERIT Matched Sample Jitter Plot
The jitter plot displays cases for the treatment and control groups prior to matching (i.e., unmatched) and for matched cases by propensity score, which is on the x-axis.

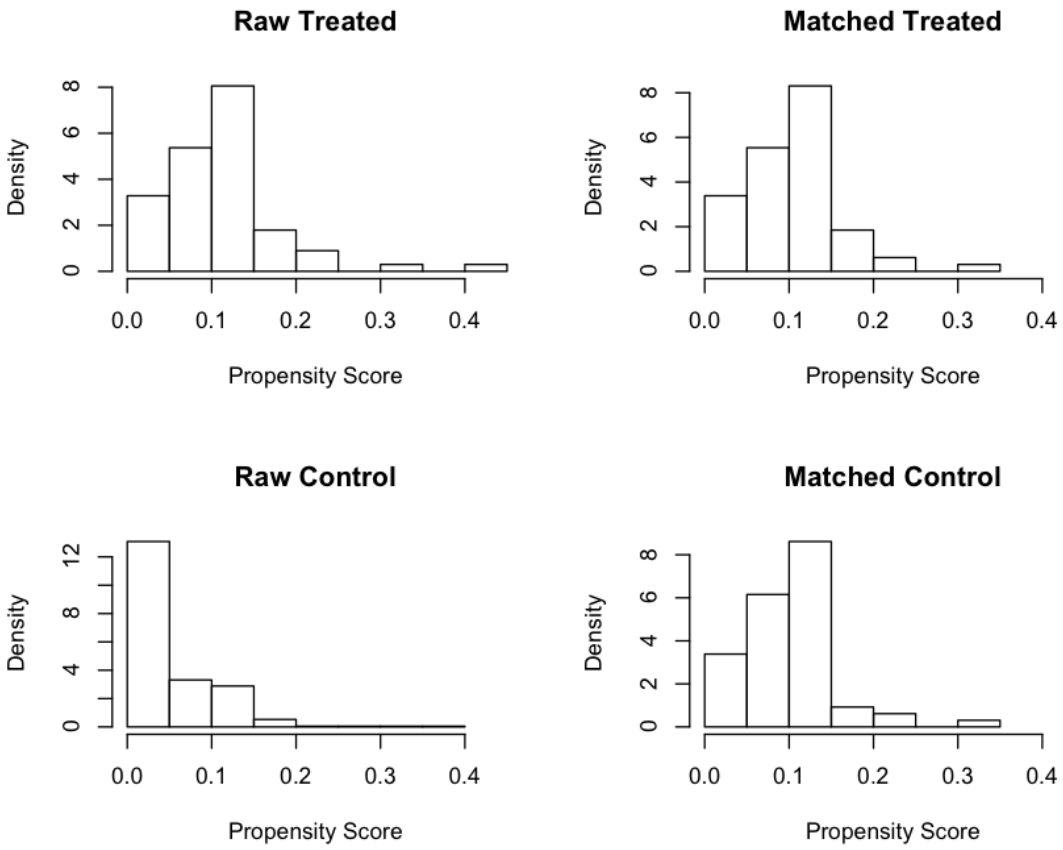

Figure 5. ERIT Matched Sample Histograms
Each graph displays a histogram for the control and treatment group by propensity score before and after matching.

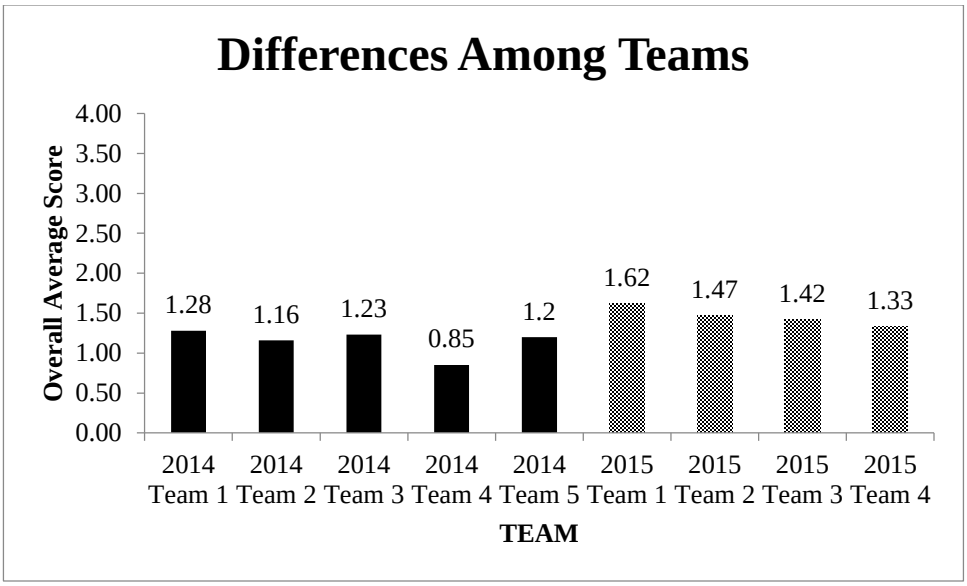

Figure 6. Differences Among Rater Teams in Overall Ethical Reasoning Essay Ratings

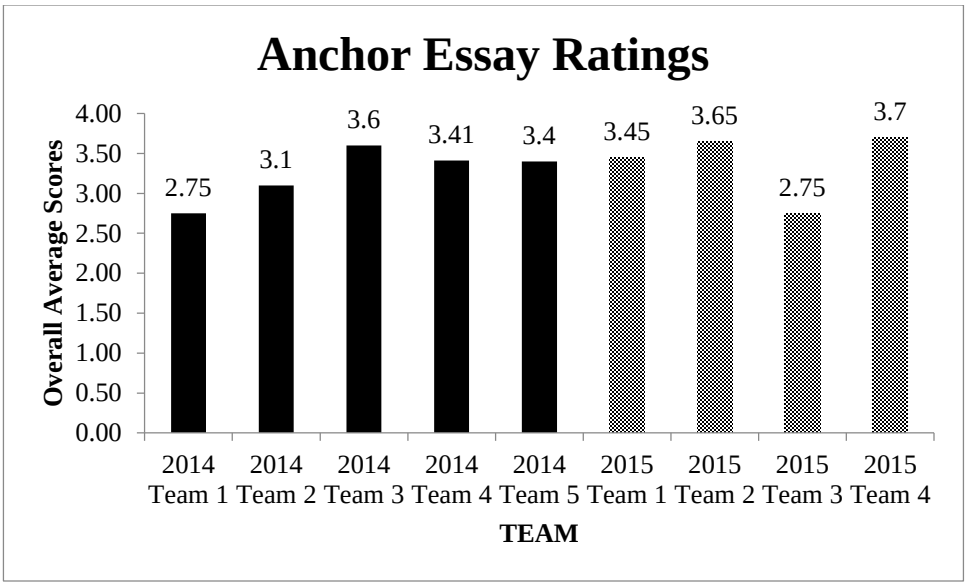

Figure 7. Anchor Essay Ratings by Rater Team

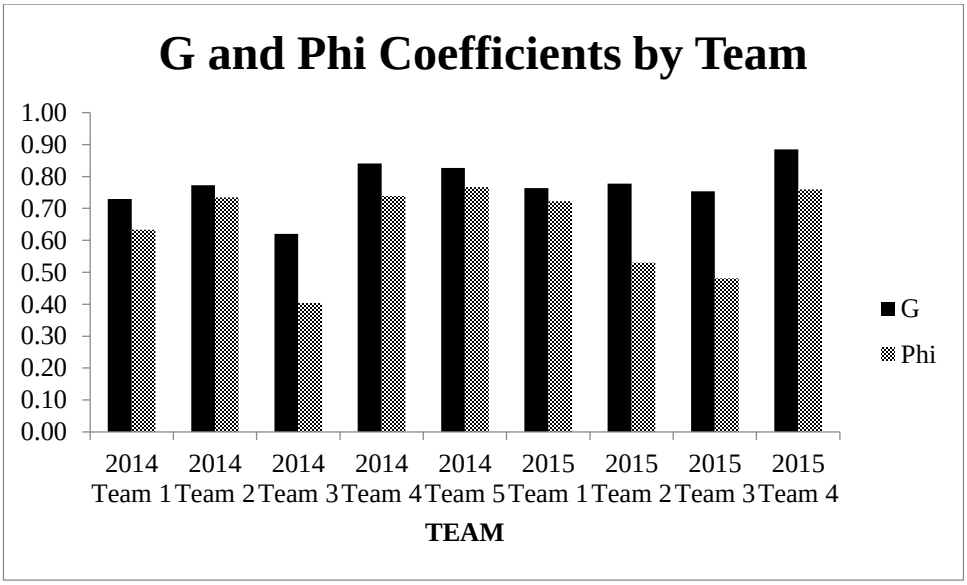

Figure 8. Essay Rating G and Phi Coefficients by Team

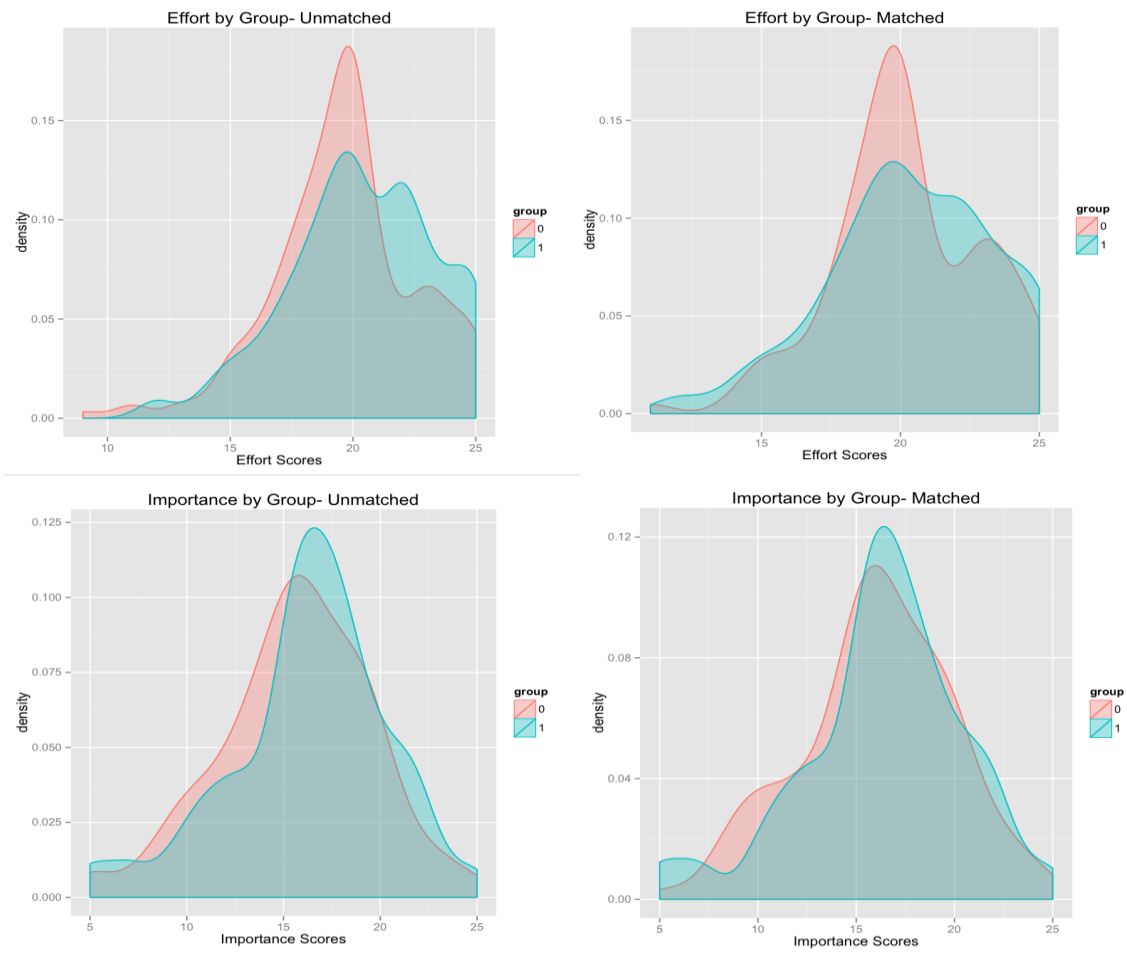

Figure 9. Effort and Importance for Treatment and Control Groups -Essays

Note. N=107 for each group. The graphs display the cumulative density function for the Effort and Importance covariates after matching.

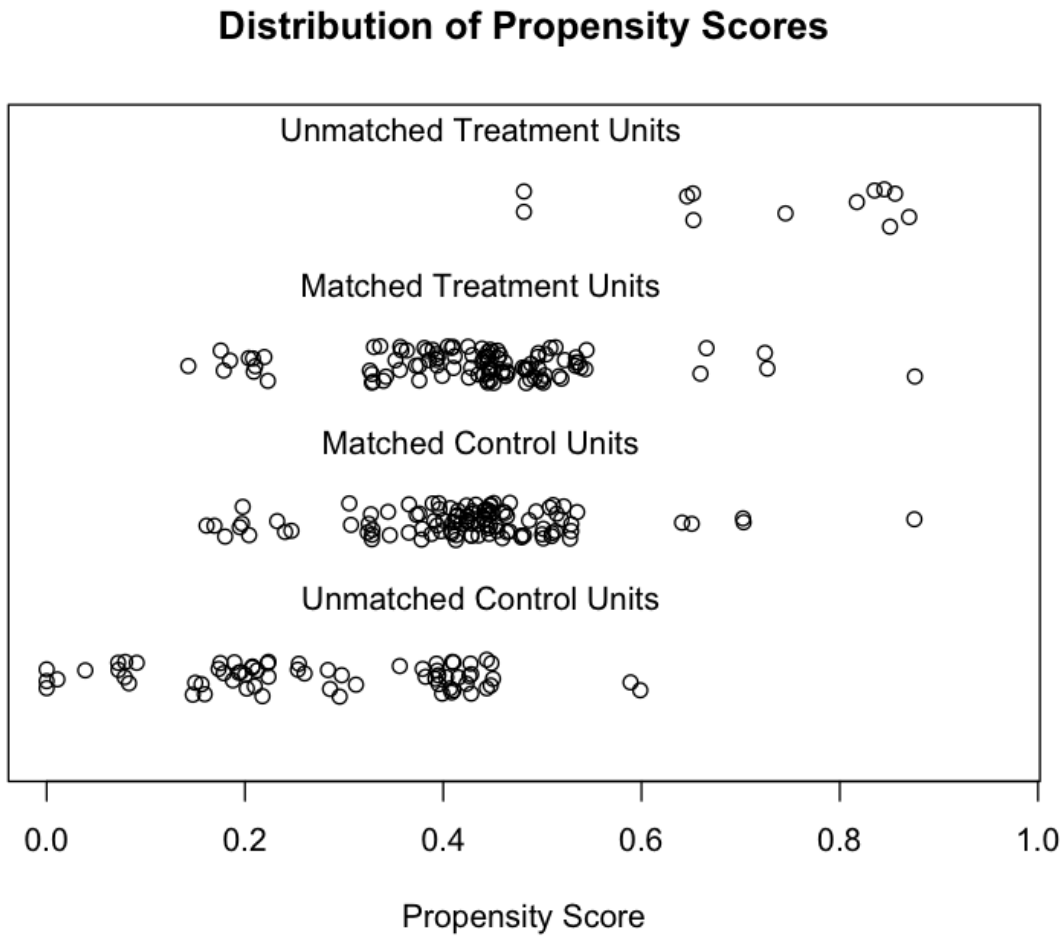

Figure 10. Essay Matched Sample Jitter Plot
The jitter plot displays cases for the treatment and control groups prior to matching (i.e., unmatched) and for matched cases by propensity score, which is on the x-axis.

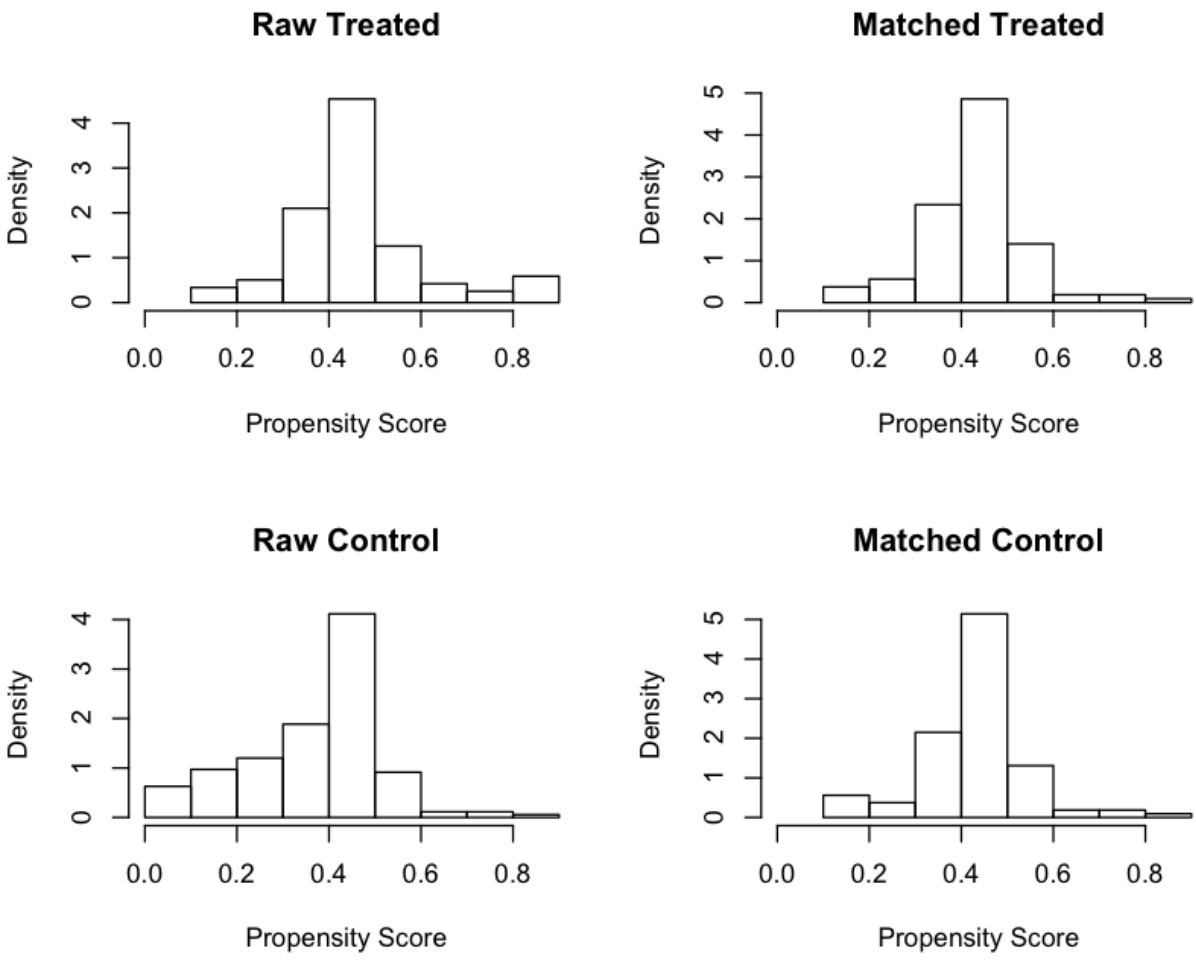

Figure 11. Essay Matched Sample Histograms
Each graph displays a histogram for the control and treatment group by propensity score before and after matching.

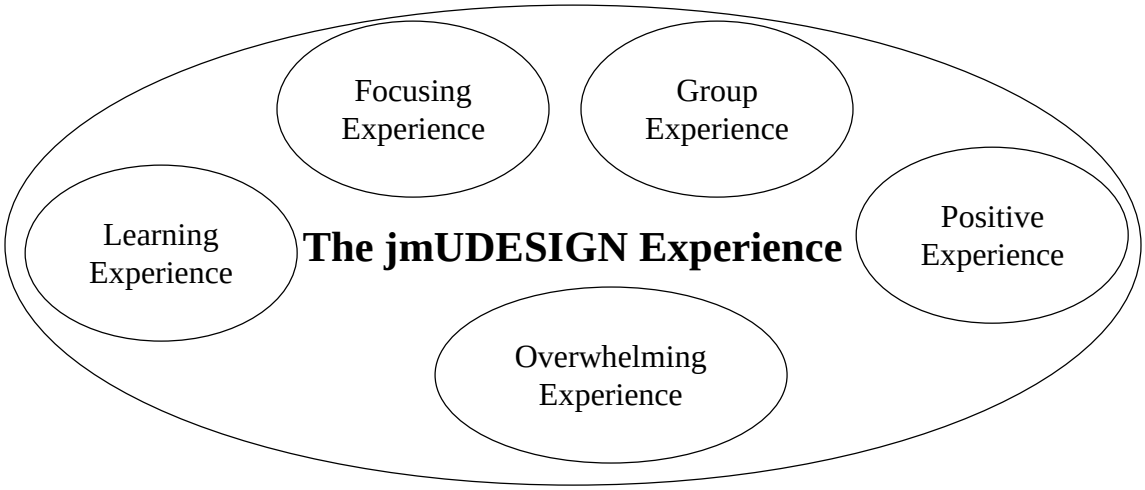

Figure 12. The jmUDESIGN Experience

Appendix A- Ethical Reasoning Rubric

James Madison University's Ethical Reasoning Rubric

Insufficient 0	Marginal 1	Good 2	Excellent 3	Extraordinary 4	Score
A. Ethical Situation: Identifying ethical issue in its context					
No reference to decision option(s).	Implicit reference to decision options AND/OR little context given regarding decision option(s).	Explicit but unorganized reference to decision option(s) and context.	Clear description of decision option(s) and context.	Meets criteria for *Excellent* AND… • Context treated with nuance • Builds tension with organization and word choice.	
B. Key Question Reference: Mentioning the 8 KQs or equivalent terms					
Reference to zero or only one key question.	Vague references to key questions OR only two key questions referenced.	References four key questions.	References six key questions.	References all eight key questions.	
C. Key Question Applicability: Describing which of the 8 KQs are applicable or not applicable to the situation and why					
No rationale provided for the applicability or inapplicability of any KQs to the ethical situation. ***SPECIAL NOTE: If author identifies fewer than three applicable KQs, then Criteria "D" and "E" can be scored no higher than (1) "Marginal"****	Provides a rationale for the applicability or inapplicability of two key questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of four key questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of six key questions to the ethical situation.	For all eight questions provides a rationale for its applicability or inapplicability to the ethical situation.	
D. Ethical Reasoning: Analyzing individual KQs					
No attempt to analyze any of the referenced key questions. ***SPECIAL NOTE: If Criterion "D" is scored a 0 or 1 then Criterion "E" can be scored no higher than (1) "Marginal"****	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.	Analysis attempted using three or more key questions. Basically accurate ascription of the key questions to the ethical situation. Account is unclear or disorganized.	Analysis attempted using three or more key questions. Accurate ascription of the key questions to the ethical situation. Account is clear and organized.	Meets criteria for *Excellent* AND… Nuanced treatment of key questions, for example: • elucidates subtle distinctions • uses analogies or metaphors • considers different issues within same key question.	
E. Ethical Reasoning: Weighing the relevant factors and deciding					
No judgment is presented OR judgment presented with no rationale.	Uses products of the analysis and provides some weighing to make a decision. Account is unclear, disorganized, or inaccurate.	Conveys weighing approach using analysis products. Provides an intelligible basis for judgment.	Meets criteria for *Good* AND…. Logically terminates in decision that will be reached.	Meets criteria for *Excellent* AND… Products of analysis weighed to make judgment compelling.	

Appendix B- The Student Opinion Survey

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

1= Strongly Disagree
2=Disagree
3=Neutral
4=Agree
5=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.

Appendix C- Summer Interview Questions

1. What products did you create during jmUDESIGN? How do they pertain to Ethical Reasoning?

2. What was it like learning about course design while simultaneously trying to infuse Ethical Reasoning into your course structure?

3. What challenges do you foresee in implementing the segment of your course that you redesigned?

4. Describe the best parts of the jmUDESIGN experience.

5. How could your experience have been improved?

6. What was the jmUDESIGN experience like for you overall?

Appendix D- Fall Interview Questions

1. Recall the jmUDESIGN experience and the product(s) you created. Have you implemented the redesigned components of your course yet? If so, please describe this experience.

2. What was the experience of teaching ethical reasoning like for you

3. What were the challenges of teaching ethical reasoning?

4. Was there anything you didn't expect that you experienced while teaching ethical reasoning?

5. Can you think of any support that would have made teaching ethical reasoning in your course easier?

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Arceneaux, K., & Nickerson, D. W. (2009). Modeling certainty with clustered data: A comparison of methods. *Political Analysis*, *17*(2), 177-190.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399-424.

Austin, P. C. (2011b). Optimal caliper widths for propensity score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150-161.

Baker, G.R., Jankowski, N.A., Provezis, S., & Kinzie, J. (2012). Using assessment results: Promising practices of institutions that do it well. *Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA)*.

Banta, T.W. & Blaich, C.F. (2011). Closing the assessment loop. *Change: the Magazine of Higher Learning (43)* 22-27.

Banta, T. W., Jones, E. A., & Black, K. E. (2010). *Designing effective assessment: Principles and profiles of good practice*. San Francisco, CA: John Wiley & Sons.

Banta, T. W., Lund, J. P., Black, K. E., & Oblander, F. W. (1996). *Assessment in practice: Putting principles to work on college campuses.* San Francisco, CA: Jossey-Bass.

Barron, K.E., Brown, A.R., Egan, T.E., Gesualdi, C.R., & Marchuk, K.A. (2008). Validity. In S. F. David & Buskist (Eds.) *21st century psychologist: A reference handbook* (pp. 55-64). Thousand Oaks, CA: SAGE.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*(1), 10-17.

Bill and Melinda Gates Foundation. (n.d.). Postsecondary Success. Retrieved February 28, 2015, from http://postsecondary.gatesfoundation.org/about/

Blaich, C.F., & Wise, K.S. (2010). Moving from assessment to institutional improvement. *New Directions for Institutional Research, 2010*(S2), 67-78.

Blaich, C.F., & Wise, K.S. (2011). From gathering to using assessment results: Lessons from the Wabash National Study. *(NILOA Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.*

Blumberg, P. (2009). *Developing learner-centered teaching: A practical guide for faculty*. San Francisco, CA: John Wiley & Sons.

Bok, D. (2013, November 11). We must prepare Ph.D students for the complicated art of teaching. *The Chronicle of Higher Education*. Retrieved from: http://chronicle.com/article/We-Must-Prepare-PhD-Students/142893/

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Carnegie Foundation for the Advancement of Teaching. (n.d.). Our Ideas. Retrieved March 8, 2015, from http://www.carnegiefoundation.org/our-ideas/

Chism, N. V. N., & Szabó, B. (1997). How faculty development programs evaluate their services. *Journal of Staff, Program, and Organization Development, 15*(2), 55-62.

Commission on Institutions of Higher Education New England Association of Schools and Colleges (2011). *Standards for accreditation: Commission on institutions of higher education New England association of schools and colleges*. Bedford, MA: Commission on Institutions of Higher Education. Retrieved from http://cihe.neasc.org/downloads/Standards/Standards_for_Accreditation_FINAL_2011.pdf

Cornell University (n.d.). Faculty Course Design Institute. *Cornell University Center for Teaching Excellence*. Retrieved July 12, 2014, from http://www.cte.cornell.edu/programs-services/faculty/course-design-institute.html

Creswell, J.W. (2013). *Qualitative inquiry and research design.* Thousand Oaks, CA: SAGE publications.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Thousand Oaks, CA: SAGE Publications.

Creswell, J.W., & Plano Clark, V.L. (2007). *Designing and conducting mixed methods research* (p. 275). Thousand Oaks, CA: SAGE publications.

Erwin, T.D. (1991). *Assessing student learning and development: A guide to the principles, goals, and methods of determining college outcomes.* San Francisco, CA: Jossey-Bass.

Fink, D.L. (2003). *Creating significant learning experience: An integrated approach to designing college courses.* San Francisco, CA: Jossey-Bass.

Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, *4*(1), 6-16.

Flannery, M. (2011, September 11). Is the US Falling behind in higher education? Retrieved February 28, 2015, from http://neatoday.org/2011/09/20/is-the-u-s-falling-behind-in-higher-education-2/

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32.

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). A Simple model for learning improvement: Weigh pig, feed pig, weigh pig. *(NILOA Occasional Paper No. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.*

Gaff, J.G., & Simpson, R.D. (1994). Faculty development in the United States. *Innovative Higher Education*, *18*(3), 167-176.

Gerstner, J. J., & Finney, S. J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Journal of Research & Practice in Assessment*, *8*, 15-29.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press, Inc.

Green, S.B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74(1), 155-167.

Hansen, E.J. (2011). *Idea-based learning: A course design process to promote conceptual understanding*. Sterling, VA: Stylus Publishing, LLC.

Hara, B. (2010, November 18). Academic freedom vs. mandated course content. *The Chronicle of Higher Education*. Retrieved from: http://chronicle.com/blogs/profhacker/academic-freedom-vs-mandated-course-content/28764

Hines, S.R. (2009). Investigating faculty development program assessment practices: What's being done and how can it be improved? *The Journal of Faculty Development, 23*(3), 5-19.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007). *MatchIt: Nonparametric preprocessing for parametric causal inference.* Software for using matching methods in R. Available at http://gking.harvard.edu/matchit/

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10-16.

Hoyt, W. T. (2010). Interrater reliability and agreement. In G.R. Hancock and R. O. Mueller (Eds.) The reviewer's guide to quantitative methods in the social sciences (141-154). New York, NY: Routledge.

Huba, M.E., & Freed, J.E. (2000). *Learner centered assessment on college campuses: Shifting the focus from teaching to learning.* Pearson.

Indiana University South Bend (n.d.). Course Design Institute. *University Center for Excellence in Teaching*. Retrieved July 12, 2014, from https://www.iusb.edu/ucet/programs/course_design_inst.php

James Madison University (n.d.). jmUDESIGN. *Institutes.* Retrieved July 12, 2014, from http://www.jmu.edu/cfi/teaching/institutes/jmudesign.shtml

James Madison University (2013, January 1). The Madison Collaborative: Ethical Reasoning in Action. Retrieved July 1, 2014, from http://www.jmu.edu/files/qep-proposal.pdf

James Madison University (2014). Wednesday Session Descriptions. *Center for Faculty Innovation*. Retrieved July 13, 2014, from http://www.jmu.edu/maysymposium/institute-descriptions.shtml

Jankowski, N. (2011, August). *Capella University: An outcomes-based institution* (NILOA Examples of Good Assessment Practice). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from http://www.learningoutcomesassessment.org/CaseStudyCapellaU.html

Jankowski, N. (2012, April). *St.Olaf: Utilization-Focused Assessment* (NILOA Examples of Good Assessment Practice). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from http://www.learningoutcomeassessment.org/documents/StOlaf.pdf

Jonson, J.L., Guetterman, T., & Thompson Jr, R.J. (2014). An integrated model of influence: Use of assessment data in higher education. *Research and Practice in Assessment, 9,* 18-30.

Kant, I. (1797). The metaphysics of morals. (Paul Guyer ed.). New York, NY: Oxford University Press, Inc. 2005.

Kinzie, J. (2012, June). *Carnegie Mellon University: Fostering assessment for improvement and teaching excellence.* (NILOA Examples of Good Assessment Practice). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from: http://www.learningoutcomesassessment.org/CaseStudyCarnegieMellon.html

Kohlberg, L. (1969). Stage and sequence: The cognitive approach to socialization. In Goslin, D.A. (Ed.), *Handbook of socialization theory and research* (p. 347-480). Chicago, IL: Rand McNally.

Kucsera, J. V., & Svinicki, M. (2010). Rigorous evaluations of faculty development programs. *The Journal of Faculty Development*, *24*(2), 5-18.

Kuh, G. D., & Ewell, P. T. (2010). The state of learning outcomes assessment in the United States. *Higher Education Management and Policy*, *22*(1), 1-20.

Kuh, G. &, Ikenberry, S. (2009, October). More than you think, less than we need: Learning outcomes assessment in American higher education. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Kuh, G.D., Jankowski, N., Ikenberry, S.O., & Kinzie, J. (2014). *Knowing What Students Know and Can Do: The Current State of Student Learning Outcomes Assessment in US Colleges and Universities*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Kvale, S. (1989). *Issues of validity in qualitative research.* Lund, Sweden: Chartwell Bratt.

Langley, G. J., Moen, R.D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide.* San Francisco, CA: Jossey-Bass.

Lee, V.S. (2010). Program types and prototypes. In Gillespie, K.J. & Robertson, D.L. (Eds.), *A guide to faculty development* (21-34). San Francisco, CA: Jossey-Bass.

Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher, 44*(1), 54-61.

Lichtman, M. (2012). *Qualitative Research in Education: A User's Guide*. Beverly Hills, CA: SAGE.

Lincoln, Y.S. & Guba, E.G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: SAGE.

Lumina Foundation. (2013). Lumina Foundation Strategic Plan. Retrieved February 28, 2015, from http://www.luminafoundation.org/files/file/2013-lumina-strategic-plan.pdf

Marvel, M.K. (1990). Improving clinical teaching skills using the parallel process model. *Family Medicine*, *23*(4), 279-284.

Merriam, S.B. (2009). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Meyer, J.P. (2010). *Reliability*. Oxford: Oxford University Press.

Middle States Commission on Higher Education (2011). *Characteristics of excellence in higher education: Requirements of affiliation and standards for accreditation* (12 ed.). Philadelphia, PA: Middle States Commission on Higher Education. Retrieved from https://www.msche.org/publications/CHX-2011-WEB.pdf

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2008). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods*, *1*(2), 13-22.

Nathan, R.G., & Smith, M.F. (1992). Students' evaluations of faculty members' teaching before and after a teacher-training workshop. *Academic Medicine,67*(2), 134-5.

Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001) Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*, 387–398.

North Central Association of Colleges and Schools: The Higher Learning Commission (2014). *The criteria for accreditation: Guiding values*. Retrieved from http://ncahlc.org/Criteria-Eligibility-and-Candidacy/guiding-values-new-criteria-for-accreditation.html

Northwest Commission on Colleges and Universities. (2010). *Standards for accreditation.* Retrieved from: http://www.nwccu.org/Pubs%20Forms%20and%20Updates/Publications/Standards%20for%20Accreditation.pdf

Nvivo for Mac (2014). Qualitative data analysis software. *QSR International Pty Ltd.* Version 10.

Onwuegbuzie, A. J., Johnson, R. B., & Collins, K. M. (2009). Call for mixed analysis: A philosophical framework for combining qualitative and quantitative approaches. *International Journal of Multiple Research Approaches*, *3*(2), 114-139.

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology*, *8*(5), 375-387.

Ouellet, M. (2010). Overview of faculty development: History and choices. In Gillespie, K.J. & Robertson, D.L. (Eds.), *A guide to faculty development* (3-20). San Francisco, CA: Jossey-Bass.

Palomba, C.A., & Banta, T.W. (1999). Assessment essentials: planning, implementing, and improving assessment in higher education. *Higher and Adult Education Series*. San Francisco, CA: Jossey-Bass, Inc.

Pepin, C.K. (2014). The dilemma of assessment in the US. In Li, Q., & Gerstl-Pepin, C. (Eds.), *Survival of the Fittest: The Shifting Contours of Higher Education in China and the United States* (pp. 73-83). Chicago, IL: Springer.

Piaget, J. (1932). *The moral judgment of the child.* Translated by Marjorié Gabain. Glencoe, IL: The Free Press.

Plano Clark, V.L., Schumacher, K., West, C., Edrington, J., Dunn, L.B., Harzstark, A., Melisko, M., Rabow, M.W., Swift, P.S., & Miaskowski, C. (2013). Practices for embedding an interpretative qualitative approach within a randomized clinical trial. *Journal of Mixed Methods Research, 7*(3), 219-242.

Proof of Concept. (n.d.). *Oxford English Dictionary.* In Oxford English Dictionary. Retrieved April 11, 2015, from http://www.oed.com/view/Entry/152578?redirectedFrom=proof+of+concept#eid28208915.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: SAGE.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Reed, T.E., Levin, J. & Malandra, G.H. (2011, September/October). Closing the assessment loop by design. *Change Magazine*, Retrieved from http://www.changemag.org/Archives/Back Issues/2011/September-October 2011/closing-the-full.html

Reindl, T. (2007). Hitting home: Quality, cost, and access challenges confronting higher education today. *Lumina Foundation for Education.*

Rutz, C., Condon, W., Iverson, E.R., Manduca, C.A., & Willett, G. (2012). Faculty professional development and student learning: what is the relationship? *Change: The Magazine of Higher Learning*, *44*(3), 40-47.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Thousand Oaks, CA: SAGE Publications.

Skeff, K.M., Stratos, G., Campbell, M., Cooke, M., & Jones III, H.W. (1986). Evaluation of the seminar method to improve clinical teaching. *Journal of General Internal Medicine, 1*(5), 315-322.

Smith, K.L. (2014). *Assessing Ethical Reasoning Skills: Initial Validity Evidence for the Ethical Reasoning Identification Test* (Master's Thesis). James Madison University.

Snijders, T.A.B. & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: SAGE.

Southern Association of Colleges and Schools: Commission on Colleges (2012). *The principles of accreditation: Foundations for quality enhancement* (5 ed.). Decatur, GA: Southern Association of Colleges and Schools: Commission on Colleges. Retrieved from http://www.sacscoc.org/pdf/2012PrinciplesOfAcreditation.pdf

Southern Association of Colleges and Schools: Commission on Colleges (n.d.). General information on the reaffirmation process. *Commission on Colleges*. Retrieved July 12, 2014 from http://www.sacscoc.org/pdf/genaccproc.asp

Stanford University (n.d.). Course Design Institute (formerly Boot Camp). *Stanford-Teaching Commons*. Retrieved July 12, 2014, from https://teachingcommons.stanford.edu/teaching-services/new-junior-faculty-assistance/course-design-institute-formerly-boot-camp.

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., Spencer, J., Gelula, M., Prideaux, D. (2006). A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME guide no. 8. *Medical Teacher, (28)*6, 497-526.

Stolorow, R. D., & Atwood, G. E. (1996). The intersubjective perspective. *Psychoanalytic Review – New York, 83,* 181-194.

Stuart, E.A. and Rubin, D.B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. Chapter 11 (pp. 155-176) in *Best Practices in Quantitative Social Science*. J. Osborne (Ed.). Thousand Oaks, CA: SAGE Publications.

Suffolk University (n.d.). Course Design Institute. *Course Design Institute*. Retrieved July 12, 2014, from http://www2.suffolk.edu/offices/52891.html

Sundre, D.L. (2007) *The Student Opinion Survey: A measure of examinee motivation.* Test manual. Retrieved March 12, 2013, from www.jmu.edu/assessment/resources/resource_file/sos_manual.pdf.

Suskie, L. (2010). *Assessing student learning: A common sense guide*. San Francisco, CA: John Wiley & Sons.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education, 58*(3), 129-151.

Tufts University (n.d.). Course Design Institute. *Center for Enhancement of Learning and Teaching*. Retrieved July 12, 2014, from http://provost.tufts.edu/celt/course-design-institute-coming-soon/

U.S. Department of Education (2013). *Accreditation in the United States*. Retrieved from website: http://www2.ed.gov/admins/finaid/accred/index.html

U.S. Department of Education: Federal Student Aid. (n.d.). *Accreditation*. Retrieved from website: http://studentaid.ed.gov/prepare-for-college/choosing-schools/consider

van Manen, M. (1990). *Researching lived experience: Human science for an action sensitive pedagogy*. Albany, NY: Suny Press.

Wabash College (2009). *Wabash national study 2006-2009*. Retrieved from http://www.liberalarts.wabash.edu/study-design/

Walvoord, B.E. (2004). *Assessment clear and simple: A practical guide for institutions, departments, and general education*. San Francisco, CA: John Wiley & Sons.

Western Association of Schools and Colleges: Senior College and University Commission (2013). *2013 handbook of accreditation*. Alameda, CA.

White House. (n.d.a). Higher Education. Retrieved November 5, 2014, from http://www.whitehouse.gov/issues/education/higher-education

White House. (n.d.b). College Scorecard. Retrieved February 28, 2015, from http://www.whitehouse.gov/issues/education/higher-education/college-score-card

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning, 28*(2), 209-220.