James Madison University
JMU Scholarly Commons

Masters Theses | The Graduate School

Spring 2014

Assessing ethical reasoning skills: Initial validity evidence for the Ethical Reasoning Identification Test

Kristen Lynn Smith
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/master201019

 Part of the Psychology Commons

Recommended Citation
Smith, Kristen Lynn, "Assessing ethical reasoning skills: Initial validity evidence for the Ethical Reasoning Identification Test" (2014). *Masters Theses*. 333.
https://commons.lib.jmu.edu/master201019/333

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

Assessing Ethical Reasoning Skills:

Initial Validity Evidence for the Ethical Reasoning Identification Test

Kristen Smith

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2014

Acknowledgments

This thesis would not have come to fruition were it not for the help, support, time, and guidance of several people, many of which have become my mentors and lifelong friends. First and foremost, I would like to extend my gratitude and thanks to my adviser, Dr. Keston Fulcher. Thank you for helping me polish my writing skills, manage this thesis research project, and think about research with a more macro perspective. I really appreciate all of your support and your willingness to give me autonomy throughout this research process. In addition, I can't thank you enough for teaching me other valuable, life lessons along the way.

I would also like to thank the other members of my committee who contributed substantially to this thesis research project. To Dr. Sara Finney, thank you for lending your measurement expertise; this project benefitted from all of your contributions. Also, thank you for patiently explaining complicated analyses to me, and supporting me throughout this research process. To Dr. William Hawk, thank you for contributing your expertise in ethical reasoning and philosophy to this thesis project. Thank you for teaching me about ethical reasoning and the Madison Collaborative. I am tremendously grateful to you for your helpful, thought-provoking insights.

I would be remiss if I did not extend my gratitude to several of my peers that helped me throughout this research process. I would like to thank Rochelle Fisher, Megan Rodgers, Jerusha Gerstner, and Bo Bashkov for all of their knowledge, support, and companionship. Many thanks to each of them for lifting me up and keeping me grounded.

Table of Contents

List of Tables

List of Figures

Abstract

Employers, policymakers, parents and other stakeholders value ethical reasoning (ER) skills. Thus, to help students actively engage in the ER process, stakeholders at James Madison University (JMU) redefined ER, implemented campus-wide ER interventions, and created the Ethical Reasoning Identification Test (ERIT-1) to measure students' ability to engage in a lower-level step of the ER process. The current study examined the factor structure and reliability of the ERIT-1. Confirmatory Factor Analysis results provided support for a unidimensional factor structure, meaning stakeholders can report and analyze total scores for the ERIT-1. ERIT-1 scores also demonstrated good reliability. Correlation analyses provided initial external validity evidence for ERIT-1 scores, indicating that the ERIT-1 and the SAT verbal reasoning test measure substantively different constructs. In addition, the ERIT-1 was sensitive to slight differences in ER training. Specifically, students experiencing a 75-minute ER intervention tended to perform better on ERIT-1 items compared to students that experienced no ER intervention. Overall, the ERIT-1 demonstrated great potential for assessing ER student learning outcomes. Future research should continue building upon this base of validity evidence. For instance, researchers should collect additional known groups validity evidence from students who received stronger doses of ER interventions.

Chapter I: Introduction

From academia to business to healthcare, society grapples with ethical dilemmas. How humanity confronts these ethical issues not only defined our past, but continues to frame our future. People's capacity to effectively address these complex problems, through ethical reasoning (ER) skills, has become the topic of national conversations. According to a national survey of employers, ER skills are "critical to a candidate's potential for career success" (Association of American Colleges and Universities, 2013, p. 1). In a similar vein, higher education has recently renewed its interest in ER skills (Dalton & Crosby, 2011). Given this new focus on ER, how would a higher education institution go about improving students' ER skills? One institution, James Madison University (JMU), used the following approach to improve students' ER abilities. First, stakeholders redefined the construct of ER in theoretical and empirical terms. Then stakeholders at JMU created a strategic intervention plan to educate students about ER, in accordance with their definition. To capture the efficacy of the intervention, stakeholders administered a robust data collection system; one part of which was a new assessment instrument. This thesis focused on collecting validity evidence for an ER assessment instrument that measures a lower level step in the ER process, as defined by JMU's Quality Enhancement Plan (QEP): *The Madison Collaborative: Ethical Reasoning in Action*.

This introduction highlights the importance of studying ER and higher education's renewed interest in this construct (Dalton & Crosby, 2011). For example, three higher education organizations (i.e., the Center for Practical and Professional Ethics, the Society for Ethics Across the Curriculum, and the Center for the Study of

Ethics) promote students' ER abilities by providing resources and research opportunities. In addition to these organizations, JMUs Madison Collaborative redefined ER and created campus-wide ER interventions to directly impact every undergraduate student. That is, to assess lower-level ER abilities the Madison Collaborative liaised with ER content experts, JMU faculty from diverse academic disciplines, and assessment consultants from the Center for Assessment and Research Studies (CARS) to develop a new instrument, the Ethical Reasoning Identification Test (ERIT). The introduction concludes by summarizing the purpose of this thesis: to better understand how to score the ERIT and collect validity evidence for ERIT scores. Collecting such evidence will contribute to JMU's efforts to cultivate students' ER abilities. Strengths and weaknesses of the ERIT will be examined using Benson's (1998) strong program of construct validation, which includes substantive, structural, and external stages.

Importance of Studying ER.

A quick scan of news headlines emphasizes the importance of cultivating ER skills. For instance, at Pennsylvania State University an assistant football coach witnessed another coach sexually assaulting a boy. He reported the incident to his direct superiors; however, they did not notify off-campus authorities. Given no action was taken, the perpetrator continued to sexually abuse numerous boys. Meanwhile, the assistant coach kept quiet about the incident; he did not reach out to off-campus authorities or other high ranking officials. The assistant coach had to grapple with a complicated ethical dilemma. He probably felt as though he had done his due diligence by reporting the incident to his direct supervisors. Not to mention, he likely recognized that reporting the incident to the authorities would jeopardize his job, and the entire

football program. However, failure to report the assault to off-campus authorities or other officials allowed the perpetrator to continue violating his victims' fundamental, personal rights. News coverage of ethical issues, like this one, situated ER at the forefront of national conversations, underscoring the importance of ER education.

Unfortunately, this particular ethical issue does not represent an isolated incident. CNN's website devotes an entire webpage to current "Ethical Issues" in our society (http://topics.cnn.com/topics/ethical_issues). Issues include congressional insider trading, corruption in state legislatures, and politician scandals. Likewise, in 2013, University of Notre Dame released its first annual "List of Emerging Ethical Issues in Science and Technology" (http://reilly.nd.edu/outreach/emerging-ethical-dilemmas-and-policy-issues-in-science-and-technology/). This list included concerns about genetic testing and eugenics; manufacturing counterfeit, life-saving pharmaceuticals; and how the use of personal data affects privacy rights. In the wake of corporate scandals such as Enron and Tyco, Albaum and Peterson (2006) emphasized the importance of "knowing the ethical perspectives of future business leaders" (p. 301).

Academia has also acknowledged the relevance of ER skills. Research demonstrated that cheating and plagiarism are a widespread "epidemic" (Alschuler & Blimling, 1995), and the number of scientific articles retracted due to fraud has increased over the past three decades (Fang, Steen, & Casadevall, 2012). One national organization, The Association of American Colleges and Universities (AAC&U) lists "Ethical Reasoning and Action" as an "Essential Learning Outcome." AAC&U emphasizes that "… students should prepare for twenty-first-century challenges by gaining Personal and Social Responsibility including Ethical Reasoning and Action anchored through active

involvement with diverse communities and real-world challenges" (The Essential Learning Outcomes, 2013). As Augustine (2013) deftly points out, "And who wants a technology-driven economy if those who drive it are not grounded in such fields as ethics?" (Globalization section, para. 7). Given expectations from employers, policy makers, and national organizations, higher education has renewed its focus on studying and developing students' ER skills (Dalton & Crosby, 2011).

Of particular interest within academe is the relationship of higher education to ethical reasoning. To this end, several studies have linked participation in college to ER development (King & Mayhew, 2002; Pascarella, & Terenzini, 1991; Rest, 1979b; Rest, Deemer, Barnett, Spickelmier, & Volker, 1986; Rest &Thoma, 1985). Rest (1979b) found that college students assessed four years after their senior year in high school showed continued gains in ER skills; however, non-college students did not. Moreover, college students gained ER skills at a *higher* rate over four years than a non-college group of participants (Rest, 1979b). As time passed, college students and non-college students became increasingly different in their ER skills; college students continued to increase while non-college students plateaued (Rest et al., 1986). This result suggested that higher education positively impacted ER abilities beyond graduation. Pascarella and Terenzini (1991) found that students advanced from lower levels to higher levels of ER while in college. That is, participation in higher education enhanced students' critical thinking skills and cognitive development, both of which are related to ER skills (Pascarella, & Terenzini, 1991). Additionally, research suggested that higher education affects students' personality characteristics and values, many of which are related to ER. For example, upperclassmen tended to have better interpersonal skills, tolerance for

differing viewpoints, and autonomy from societal impositions than freshman (Pascarella, & Terenzini, 1991). These findings implied that higher education participation promotes ER abilities.

ER Interventions.

Research suggested participating in higher education promotes ER skills (Pascarella, & Terenzini, 1991; Rest, 1979b; Rest et al., 1986; Rest & Thoma, 1985); however, these studies did not include colleges that implemented targeted, campus-wide ER interventions. Nevertheless, Kohlberg (1977) noted that ER abilities require *effortful* development. It would seem logical that strategic ER interventions that involve *all* students at an institution should yield more dramatic ER gains than simply participating in higher education generically.

Although many higher education institutions provide resources to support ER development, these initiatives rarely impact all students. Instead, institutions typically offer resources that affect a particular group of faculty or students. For instance, the Center for Practical and Professional Ethics, the Society for Ethics Across the Curriculum, and the Center for the Study of Ethics are three organizations committed to promoting ER across different academic disciplines. They support scholarship on ethics and the teaching of ethics. The Society for Ethics Across the Curriculum also publishes a peer-reviewed academic journal, *Teaching Ethics*, which examines ethical issues across academic disciplines.

These organizations promote ER skills by providing opportunities for scholars to conduct and present research about ethics. The Center for Practical and Professional Ethics hosts an annual ethics symposium and offers workshops for faculty members to

enhance the role of ethics education in their curricula. The Society for Ethics Across the Curriculum sponsors an international conference about ethical inquiry and teaching across the curriculum. Lastly, the Center for the Study of Ethics sponsors public forums and extra-curricular student scholarship in ethics. These organizations offer resources that promote ER, yet it is not part of their mission to directly impact *every* student.

The Madison Collaborative: Ethical Reasoning in Action.

JMU administrators and stakeholders endeavored to improve all undergraduates' ER skills. To achieve this goal, the Madison Collaborative designed campus-wide ER interventions, and created innovative tools to assess their effect. What follows is an explanation of their theoretical and empirical definitions of ER, a brief description of the planned interventions, and an introduction to an instrument created to assess students' lower level ER skills.

First, the Madison Collaborative articulated a *conceptualization of* ER guided by an ethics specialist; this conceptualization differs from those endorsed by Piaget (1965); Kohlberg (1969, 1977, 1984); and Rest, Cooper, Coder, Masanz, and Anderson (1974). Instead of schemas or stages, ER was defined as a process that consists of open-ended inquiries; these inquiries focus on multiple ethical considerations, which is known as a casuistry approach (William Hawk, personal communication, June 20, 2013). Specifically, the multiple ethical considerations are framed as eight Key Questions (KQs):

- Empathy – How would I respond if I cared deeply about those involved?
- Character – What actions will help me become my ideal self?
- Fairness – How can I act equitably and balance all interests?

- Liberty – What principles of freedom and personal autonomy apply?
- Rights – What rights (e.g., innate, legal, social) apply?
- Responsibilities – What duties and obligations apply?
- Outcomes – What are the short-term and long-term outcomes of possible actions?
- Authority – What do legitimate authorities (e.g., experts, law, my god[s]) expect of me?

Students must consider which KQs are relevant to a particular ethical situation then weigh and balance those KQs to inform a choice or decision leading to a course of action. Given context was essential to the ER processes, the Madison Collaborative incorporated three areas of application. Students should apply the eight KQs to *Personal*, *Professional*, and *Civic* areas. Including three application areas expanded the breadth of the theoretical domain and supported the Madison Collaborative's emphasis on "ER in action."

The theoretical definition of ER was operationalized into student learning outcomes (SLOs) or, in other words, what the Madison Collaborative expected students to learn, think, or do. The five cognitive SLOS are as follows:

1. Students will be able to state, from memory, all Eight KQs.
2. When given a specific decision and rationale on an ethical issue or dilemma, students will correctly identify the KQ most consistent with the decision and rationale.
3. Given a specific scenario, students will identify appropriate considerations for each of the Eight KQs.

4. For a specific ethical situation or dilemma, students will evaluate courses of action by applying (weighing and, if necessary, balancing) the considerations raised by KQs.
5. Students will apply SLO 4 to their own personal, professional, and civic ethical cases.

Progression through the SLOs is hierarchical such that a student must achieve an earlier outcome before achieving subsequent outcomes. Notice how the SLOs increase in complexity beginning with memorization and culminating in real-world application. Achieving each of these learning outcomes should improve students' ER abilities. That is, a student would improve their ER abilities if they only learned how to identify the KQ most consistent with a given rationale (i.e., SLO 2). However, the student would further refine their ER skills and gain ER expertise as they proceeded through the hierarchy to SLO 5.

After defining ER through the Key Question Framework, and establishing measureable SLOs, the Madison Collaborative designed a comprehensive ER intervention plan (See Figure 1). The purpose of the interventions was to "transform JMU into a community recognized for producing contemplative, engaged citizens who apply ethical reasoning to confront the challenges of the world" (The Madison Collaborative: Ethical Reasoning in Action, 2013, p. 22). Specifically, the interventions were intended to improve ER skills by teaching faculty and students how to approach ethical issues using the KQ model. Not to mention, the interventions were also intended "to ensure students achieve the SLOs" (The Madison Collaborative: Ethical Reasoning in Action, 2013, p. 33).

The Madison Collaborative's interventions in chronological order of implementation, included:

1) a 75-minute session during August Orientation for freshmen entitled "It's Complicated: Ethical Reasoning in Action;"
2) an online course for freshmen spanning eight months of the academic year;
3) ER programming in residence halls;
4) and curricular interventions including coverage of the 8 KQ Framework in General Education, Major-specific, and Honors courses (The Madison Collaborative: Ethical Reasoning in Action, 2013, p. 32-33).

The campus-wide, comprehensive intervention plan should help students achieve the Madison Collaborative SLOs; thus, improving their ER abilities. Recall, the SLOs were written to measure ER as an inquiry process that emphasizes a set of common ethical considerations defined by the eight KQs. Yet, no existing measures use the KQ model. Due to the misalignment between existing measures and the eight KQs, the Madison Collaborative developed several instruments that were specifically aligned to the eight KQs. Among them was the ERIT, a new measure to assess a lower-level step in the ER process (i.e., SLOs 2 and 3) (See Figure 2).

The ERIT is a cognitive, multiple-choice test; it can be administered to a large cohort of students and graded using scantron forms. The ERIT presents students with ethical scenarios and asks them to match each scenario with the most appropriate KQ. This requires students to identify the appropriate considerations associated with each KQ. Each item on the ERIT consists of eight response options, one for each KQ. In addition, each item maps onto one of the eight KQs and one of three application areas: *Personal*,

Professional, and *Civic* (The Madison Collaborative: Ethical Reasoning in Action, 2013, p.66). For example, the following ERIT item aligns with the Outcomes KQ and the Civic application area because it highlights the long-term outcomes of proposing a clean air bill, and it approaches the issue of air quality within a civic context:

> *Proposing a clean air bill would be costly both in terms of money and in political capital. Nevertheless, the senator believed that in the long term, most Americans would be healthier with better air quality.*

Any instrument used to assess Madison Collaborative initiatives must define ER using the eight KQ framework. The Madison Collaborative's large-scale ER intervention also required assessment instruments that could be administered to a large cohort of students and easily be scored. Lastly, the Madison Collaborative needed an instrument that could demonstrate growth in students' ER abilities. The regional accrediting body, the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC), required evidence of the effectiveness of campus-wide interventions; thus, the instrument had to be sensitive to changes in students' ER abilities from pre- to post-intervention.

Statement of Purpose.

The ERIT is a newly developed instrument designed to measure a lower-level step in the ER process (i.e., SLOs 2 and 3). To make inferences from students' ERIT scores, JMU stakeholders need validity evidence. Therein lies the purpose of this thesis: to understand how to score the ERIT and collect validity evidence for ERIT scores through Confirmatory Factor Analysis (CFA) and correlation analyses. As detailed in chapter two, the four research questions will relate to 1) the factor structure and reliability of the

ERIT scores; 2) the relationship between scores on the ERIT and verbal proficiency scores; 3) the relationship between scores on the ERIT and scores on another Madison Collaborative ER measure; and lastly, 4) the comparison of two different groups' performance on ERIT items.

Chapter II: Ethical Reasoning Literature Review

Although ER has recently become the focus of national conversations (Association of American Colleges and Universities, 2013; Dalton & Crosby, 2011; Fang et al., 2012; Treviño & Nelson, 2011), researchers have studied ER for centuries. The following review provides background information about the construct of ER. Specifically, the review relates ER to other constructs, investigates various ER theories, and examines measures of ER. Such information is crucial to the theoretical and empirical definition of ER used to create the Ethical Reasoning Identification Test (ERIT). The review also describes measure creation, revision, and existing reliability evidence. Benson's (1998) construct validation framework was used to evaluate the ERIT, resulting in four research questions.

ER and its Relation to Other Constructs

Recall, the Madison Collaborative defined ER as a "process," influenced by multiple philosophies. To understand what ER is, consider how it relates to constructs such as moral reasoning, civic mindedness, and openness to experience.

Ethical and Moral reasoning. ER and moral reasoning are closely linked in theory. The constructs differ in nuanced, practical ways. For instance, the words "ethics" and "morals" can have different meanings because they come from two distinct language traditions. "Morals" originates from the Latin "mores," which is a descriptor used to convey how people act (William Hawk, personal communication, June 20, 2013). "Ethics," on the other hand, comes from the Greek "ethikos" meaning a system of normative standards (William Hawk, personal communication, June 20, 2013). Thus, moral reasoning can be conceptualized as an empirical behavior, or how society *actually*

reasons. However, ER refers to how society or persons *should* reason. For example, society often has a "code of ethics" rather than a "code of morals" because ethics represent *normative standards* of reasoning or the kind of reasoning society aspires to embody.

From an academic standpoint, ER is the broader concept that subsumes moral reasoning. That is, "…developing a person's ER skills can subsequently improve their moral reasoning skills. One should attend more closely to the appropriate normative standards of reasoning so that their actual practice would improve" (William Hawk, personal communication, March 13, 2013). Furthermore, Keller (2010) defined "the practice of ethics" as "applied methods of rational inquiry to moral problems" (p. 12). In other words, ER is a means by which to investigate or consider moral issues.

In previous research, the participants' stage of development was likely a factor in determining whether the term ER or moral reasoning was used. Traditionally, researchers used the term moral reasoning instead of ER because they focused on reasoning in children or adolescents (Kohlberg, 1969, 1977, 1984; Piaget, 1965; Rest et al., 1974; Rest et al., 1986; Rest & Thoma, 1985). Furthermore, previous research on ER was empirically driven; researchers investigated how people or children *actually* reasoned (William Hawk, personal communication, November 19, 2013). A key distinction between ER and moral reasoning is that ER assumes a higher level of cognitive development than moral reasoning (William Hawk, personal communication, June 20, 2013). Given moral reasoning refers to behaviors, whereas ER is a set of normative standards, children or adolescents may not be developmentally mature enough to engage in ER processes. Thus, it is appropriate to study moral reasoning in a younger population

because it is questionable whether ER is possible at that developmental stage. During childhood and early teenage years, it is unlikely that individuals have established or adopted standards of how they *ought* to reason. When studying a group of older individuals, say college-aged students or adults, it is appropriate to assess ER abilities.

Earlier researchers often used the term moral reasoning, whereas more contemporary researchers tend to use the term ER. Some researchers, like Psychologist Lawrence Kohlberg (1977), even used these terms interchangeably. Although ER and moral reasoning can be differentiated via philosophical nuances, this review treats them *synonymously* from this point forward.

Civic mindedness. Recall, the Madison Collaborative focused on "ER in action*."* The conceptualization of "ER in action" is related to civic mindedness because students are expected to *apply* their ER skills in civic contexts (The Madison Collaborative: Ethical Reasoning in Action, 2013, p. 64). ER is related to civic mindedness in multiple ways. First, being civic-minded involves action; a civic-minded professional collaborates with others to fulfill community needs. Similarly, as a result of Madison Collaborative interventions, students must use their ER skills in civic situations. Second, Hatcher (2008) defined the concept of a "civic mindedness professional" as "one who is skillfully trained through formal education, with the ethical disposition as a social trustee of knowledge, and the capacity to work with others in a democratic way to achieve public goods" (p. 21). Hatcher's definition suggested that "ethical dispositions" are characteristic of a civic-minded professional. Moreover, his definition involved achieving "public goods." This is akin to the Madison Collaborative's expectation that

students will apply their ER skills to civic situations. Conceivably, "civic situations" would include ways "to achieve public goods."

Another source of overlap between these constructs is evidenced through instrumentation. Hatcher (2008) created the Civic Minded Professional (CMP) scale to measure the concept of "civic-minded professional." Items on the CMP scale are delineated into five subscales: Voluntary action, identity and calling, citizenship, social trustee, and consensus building. Items from the social trustee subscale such as, "I think that students have a civic responsibility to improve society by serving others" and "The knowledge I have gained throughout my life should be used to help serve others" are especially applicable to the concept of "ER in action."

Furthermore, the nature of civic-minded action is often altruistic, and altruism is related to moral reasoning (Andreason, 1976; Rubin & Schneider, 1973; Schwartz, Feldman, Brown, & Heingartner, 1969). Augusto Blasi (1980) reviewed 19 studies that attempted to relate children and adolescents' moral reasoning development to altruistic behaviors such as sharing toys or candy, helping a classmate, volunteering, or helping a researcher. Of the 19 studies, 11 found a clear relationship between moral reasoning and altruistic behavior (Blasi, 1980). The voluntary action and social trustee subscales of Hatcher's CMP scale, discussed previously, contain items that illustrate altruistic behaviors. For example, items such as "I am very willing to volunteer my time to participate in community service" and "It is important for students to give a portion of their time to community or voluntary service" elicit altruistic concepts of volunteerism and giving back to the community. ER is related to civic mindedness because being a "civic-minded professional" involves altruistic attitudes and actions.

Openness to experiences. ER is also associated with personality factors, particularly "openness to experience" (Harris, Mussen, & Rutherford, 1976). Students that have more of what Costa and McCrae (1992) defined as "openness to experience" could be more prone to exploring the different ethical considerations that constitute the KQs. That is, students that are more open, flexible, intellectually curious, and imaginative could be better equipped to consider the open-ended inquires that define the ER process compared to students that have less "openness." Also, the positive relationship between openness and moral reasoning development implies that students who are more open to experiences might be the same students that excel at ER (Dolinger & LaMartina, 1998).

Harris and colleagues (1976) found that advanced stages of moral reasoning were related to characteristics of openness to experience. Teenage boys at advanced stages of moral reasoning, as measured by Kohlberg's Moral Judgment Interview, considered themselves to be "capable of coping with new situations as they arise" (Harris et al., 1976, p. 131). In a sample of college-aged individuals, men at the highest "principled" stage of moral reasoning described themselves as being "especially open to experience (curious, sympathetic, responsive, and not reserved)" (Haan, Smith, & Block, 1968, p. 192).

Scores from various measures of ER, such as Kohlberg's Moral Judgment Interview and the Defining Issues Test (DIT), also correlate with openness. Dolinger and LaMartina found that moral reasoning, as measured by P scores on the DIT, were significantly, positively correlated with openness, $r = .30$, $p < .001$ (1998). Interestingly, no other big five personality factors measured by the NEO Personality Inventory significantly correlated with DIT scores. Moreover, openness $_{(x1)}$ was significantly

correlated with DIT scores $_{(y1)}$ after partialling vocabulary $_{(x2)}$, college academic performance $_{(x3)}$, and enjoyment of reading $_{(x4)}$ out of both openness and DIT scores, $r_{x1,y1.x2,x3,x4} = .18$, $p = .05$. In other words, openness and DIT scores are significantly correlated, *independent of* vocabulary, college academic performance, and enjoyment of reading. These findings suggested that openness is related to moral reasoning or ER development (Dolinger & LaMartina, 1998).

Investigating Various Theories of ER- A Casuistry Approach

The Madison Collaborative used various theories spanning centuries of philosophical inquiry to define the ER process. Although there are numerous theories of ER, theorists often use only one theory to address an ethical situation. Alternatively, some choose to handle ethical issues without subscribing to any ER theory. Yet another approach is to use casuistry, a method that posits that no one theory is adequate; however, integrating different theories leads to better reasoning outcomes (William Hawk, personal communication, June 20, 2013). The Madison Collaborative defined the ER process using this casuistry approach, which combines ideas from the following philosophical perspectives: John Stuart Mill's Utilitarian theory, Kant's natural duties and obligations, Rawls' justice as fairness, Kohlberg's role of authority, Gilligan's role of empathy, and Aristotle's virtuous self. Each philosophical perspective informed at least one of the eight KQs used to define the ER process. For instance, the work of philosopher John Stuart Mill contributed to the "Outcomes" KQ. Thus, the ER process, as defined by the Madison Collaborative, does not emanate from one theory; instead, it is defined as a process that is influenced by numerous philosophical perspectives.

Outcomes. In the context of utilitarianism, Mill postulated that happiness is the ultimate end goal for humanity; yet, this happiness isn't necessarily self-focused (Skorupski, 2005). Mill's ideal conception of happiness was to promote the greatest possible good for the most members of society, while simultaneously reducing suffering. In the words of Cahn & Markie (2002), "the ultimate end… is an existence exempt as far as possible from pain, and as rich as possible in enjoyments, both in point of quantity and quality" (p. 349). Moreover, the standard of conduct, or the *normative standard*, should be equal happiness for everyone (Skorupski, 2005). Mill (2003) highlighted acting in a way that would benefit others because he believed that an individual should act in the best interest of their society as a whole, which in turn provides protection for the individual. Mill (2003) states that:

> …and to perform certain acts of individual beneficence, such as saving a fellow-creature's life, or interposing to protect the defenseless against ill usage, things which whenever it is obviously a man's duty to do, he may rightfully be made responsible to society for not doing. (p. 81)

Not only should individuals generously help their fellow man, they should be held responsible if they fail to do so.

Mill's endeavor to practice ethics by maximizing happiness for all seems laudable enough; however, real-world dilemmas involve situations in which utility theory does not offer a definite solution. Take for instance Phillipa Foot's trolley car problem (1967): The driver of a runaway cable car must decide if he will steer the car one of two directions. If he steers one direction, five men working on the track ahead will perish; if he steers the

other way, the car will kill one man working on the tracks. In accordance with utility theory, the driver must act in favor of the greater good, opting to kill one rather than five.

Although the Utilitarian response to the trolley car problem proposed by Foot seems adequate, consider Judith Jarvis Thomson's (1985) medical adaptation of the trolley car problem: A surgeon has five transplant patients that need a different organ to survive; however, there are no organs currently available for transplant. If the patients do not receive new organs today, they will die. In strolls a healthy young traveler that has the same blood type as the five sick patients. The surgeon could choose to harvest the organs from the young patient, taking one life to save five. It seems that the Utilitarian solution to Thomson's "trolley car" is not quite as convincing as it is for Foot's problem. Herein denotes a potential flaw of Utilitarian theory. As espoused by Bernard Williams, utilitarian theory is *inadequate* because it does not give enough consideration to the philosophy of real moral problems, or the concept of happiness at the individual level (Smart & Williams, 1973). Williams also faulted Utilitarian theory for its inattention to human integrity. Through his hypothetical examples, he suggested that utility theory fails to consider the idea that individuals are *personally* responsible for their actions, *not* for the actions of others. Williams urged utilitarian philosophers to give further consideration to the distinction between two different individuals' happiness and motives (Cahn & Markie, 2002).

While Utilitarian ideals alone may be inadequate for practicing ethics in the real world, integrating utility theory with others can result in better reasoning outcomes. For instance, when faced with an ethical dilemma, utility theory would encourage society to consider "*What are the short-term and long-term outcomes of possible actions*?" and

"*What will result in the greatest good for the greatest number of people*?" Utilitarian ideals are useful for addressing ethical questions regarding outcomes.

Rights, liberty, and responsibilities. In addition to Utilitarian theory, Immanuel Kant's theory of natural duties and voluntary obligations was incorporated into the definition of ER. Many of Kant's philosophies were related to ethical situations that invoked questions of *rights*, *liberty*, and *responsibilities*. Kant underscored natural, inborn duties that individuals must fulfill to themselves and to their fellow human beings.

In the *Metaphysics of Morals*, Kant emphasized safeguarding rights to freedom for people as individuals and for people as members of a larger society (Guyer, 2004). According to Kant's (1797) universal principle of right, "Any action is right if it can coexist with everyone's freedom in accordance with a universal law, or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law" (p. 230). Kant also advocated for every human being's innate, "private" rights to property. According to Kant (1797), rights to property in the form of objects were not inherent because one must acquire property. However, Kant (1797) explained that rights still apply because claiming something for one's own requires mutual consent from others who, in turn, must give up their freedom to claim that particular property. The right to property was established as a societal right rather than a natural right (Guyer, 2004). Individuals were not born with rights to property; instead, property rights were socially construed and maintained by a governing body. Kant proposed that property rights were subject to regulations from governing bodies because when one person claims a property they are directly limiting the freedom of another person to claim that property. In contrast, an individual's personal beliefs do not directly limit the freedom of others,

thus personal beliefs were *not* subject to regulations from other people. In addition to these private rights, Kant (1797) made an argument for "public" rights, such as the right of the governing body to punish wrongdoers.

Concerning duties or responsibilities, Kant (1797) defined two types: duties of right and duties of virtue. Duties of right were related to Kant's "universal principle of right," stated previously. These duties existed to protect the freedom of every individual person. More specifically, duties of right focused on individuals' relationships with one another, and they could be imposed by external governing bodies (Denis, 2012). Every individual person had certain "duties of right" or responsibilities to their fellow human beings that they were expected to fulfill. Kant's duties of virtue, on the other hand, concerned individuals' inner freedoms and human nature. Thus, these duties are strictly internal to each individual; duties of virtue can be thought of as "duties to oneself." These duties included avoiding lying, drunkenness, suicide, gluttony, envy, and ungratefulness, while promoting sympathy and charity toward others (Denis, 2012).

Some faulted Kant's ideas about rights, liberty, and responsibilities. A predecessor of Kant, David Hume, had a more empirically based outlook that stood in opposition to Kant's philosophies. Hume focused on answering philosophical questions using an experimental approach that relied on naturally existing phenomena (Denis, 2012). Using a naturalistic point of view, Hume considered morality to be a naturally occurring human phenomenon that was independent of spiritual entities or religion (Denis, 2012). He asserted that in order to fulfill a duty, a person must desire to do the "right" thing first. Human traits and inclinations *cause* action or fulfillment of duties; it wasn't enough to just have an innate knowledge of these duties (Denis, 2012). That is,

Kant suggested human beings fulfilled duties, in part, due to natural or innate mechanisms. Hume, however, thought the fulfillment of duties was caused by occurrences in the physical world. Kant tended to take an innate, "nature" approach, whereas Hume upheld a more evidence-based, "nurture" approach.

Regardless of the "nature or nurture" controversy, Kant's ideas of natural duties and obligations were incorporated with other theories to define ER. To practice ER, Kant would urge individuals to consider "*What rights (e.g., innate, legal, social) apply*?"; "*What principles of freedom and personal autonomy apply*?"; and "*What duties and obligations apply*?" Kantian theories can be used to address ethical dilemmas involving issues of rights, liberty, and responsibilities.

Fairness. Another consideration that defined the ER process was "fairness" or "justice." John Rawls' work, *A Theory of Justice*, supported the inclusion of "Fairness" in the definition of ER. Influenced by Utilitarian and Kantian philosophies, Rawls (1971) presented the idea of "justice as fairness." He asked his audience to create principles of justice given no information about their economic status, abilities, intelligence, health, physical strength, etc. (Rawls, 1971). This thought exercise established a system of justice behind a "veil of ignorance." Rawls (1971) reasoned that lack of details about one's own characteristics would yield principles of justice that were fairer to every person, regardless of their actual position in society. In essence, he thought that having people create a system of justice shrouded by the veil of ignorance would situate everyone on a more equal playing field. The veil of ignorance prevented a person from privileging their own interests, thus diminishing a major barrier to achieving justice.

Based on his hypothetical thought exercise, Rawls presented two principles of justice that he thought rational persons would agree to behind the veil of ignorance. These principles of justice "affirmed the priority of equal basic liberties over other political concerns, and required fair opportunities for all citizens, directing that inequalities in wealth and social positions maximally benefited the least advantaged" (Freeman, 2002). The two principles provide a basic societal structure; they control the diffusion of duties and rights (Cahn & Markie, 2002). The primary principle, often called the "Liberty Principle," states that "each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others" (Rawls, 1971, p. 60). According to Rawls' (1971) secondary principle, "Social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all" (p. 60). The rights addressed in the Liberty Principle include more fundamental human rights and they should never be compromised to meet the second principle. The second principle is referred to as the "Difference Principle." This establishes the presumption that people should be entitled to equality, unless there is an inequality. When possible, wealth and resources should be distributed in a way that produces better outcomes for *everyone*. Said another way, people should be equal; if any inequality should occur, it must be to the advantage of the worst-off member of society (Freeman, 2002). According to Rawls, an inequality was ethically permissible *only* if it benefitted the least advantaged members of society (William Hawk, personal communication, November 19, 2013).

The definition of ER incorporates notions of "justice as fairness" proposed by Rawls. His concepts can be used to make decisions about ethical issues regarding

fairness. When disputes or inequalities arise in society, the question of fairness can help adjudicate conflicting perspectives and settle these disputes (William Hawk, personal communication, November 19, 2013). Using Rawls' theory, one should consider "*How can I act equitably and balance all interests?"* or *"What action will best achieve justice?"* or *"How is reciprocity best achieved?"*

Authority. The eight KQ framework recognizes and incorporates the role of authority in ER. For some, the answer to ethical situations often derives from a higher authority such as a religious or spiritual entity. Jean Piaget's research described the development of moral reasoning in children. He found that infants have no awareness of the world around them, outside of their own experiences (Piaget, 1932). In fact, until children were old enough to develop what Piaget called a "theory of mind," they thought other people automatically had the same understandings, beliefs, and intentions as they did. Thus, infants and young children often acted to satisfy their own needs or desires, without regard to authority figures external to their own conscientiousness. Piaget's research informed the subsequent work of Lawrence Kohlberg regarding the role of authority in moral reasoning development.

Kohlberg (1969) established what has become one of the most widely studied theories of moral development. According to Kohlberg, moral reasoning development occurs in six hierarchical stages, where each stage represents more sophisticated moral reasoning skills. He claimed that stage two reasoning yields better solutions to moral dilemmas than stage one reasoning, stage three reasoning better solutions than stage two, etc. He found that moral reasoning abilities do not develop due to maturation alone; progression from one stage to the next requires effortful development (Kohlberg, 1977).

Many of Kohlberg's stages are defined in terms of an individual's sentiment toward authority**.** As detailed in Kohlberg's (1969) *Stage and Sequence: The Cognitive Approach to Socialization,* the most basic, "preconventional" level of moral reasoning entail stages one and two. These stages are characterized by an unquestioning obedience to authority, and are dominated by egocentric tendencies to fulfill selfish needs. Basic punishments and rewards dictate actions. Stages three and four represent a transition into more "conventional" reasoning where egocentric tendencies give way to the desire to conform to societal expectations, uphold authority figures, and care about the perspective of others (Kohlberg, 1969). "Conventional" moral reasoning also involves preserving social order and personal relationships (Davison & Robbins, 1978). The final two stages represent "Postconventional" levels of reasoning. The abilities manifested at stages five and six involve abstract, ethical principles such as "the Golden Rule;" these stages represent Kohlberg's (1977; 1984) most advanced conceptualization of moral reasoning. Stage 5 and 6 define moral reasoning in a way that can be completely separated from authority groups or the individual's identification with authority groups (Kohlberg, 1977). That is, individuals start to develop a personal sense of what is right and wrong rather than submitting blindly to authority. Here, more abstract ideas of justice and equality become priority. Individuals at "postconventional" levels understand that what is "just" or "right" depends on culture, time, and personal values (Kohlberg, 1969).

Recall, the process of ER consists of open-ended inquires that emphasize various ethical considerations, including the role of authority. The notion of considering authority figures' expectations when engaging in the ER process is supported by Piaget and Kohlberg's theories. Yet, to logically discern what constitutes a "legitimate" authority

requires one to achieve at least a "Conventional" level of moral reasoning (William Hawk, personal communication, November 19, 2013). Considering what constitutes a "legitimate" authority and acknowledging the role of "legitimate" authority as part of the eight KQ framework is especially useful when ethical struggles elicit questions such as "*What do legitimate authorities (e.g., experts, law, my god[s]) expect of me?"* .

Empathy. Carol Gilligan's work concerning empathy and moral devolvement also supported the Madison Collaborative's definition of ER. Much of Gilligan's work investigated the differences between female and male moral reasoning development (Tong, 1998). In particular, she emphasized the role of empathy**,** along with other characteristically female traits, in the development of moral reasoning skills. Gilligan criticized Lawrence Kohlberg's theory of moral development because it was created based on sampling only male subjects (Hoose, 1998). She asserted that Kohlberg's stages of moral development failed to incorporate a perspective of empathy, in part, because he did not study moral development in females (Gilligan, 1982). Furthermore, Gilligan (1982) claimed that women often never reached Kohlberg's highest, "principled" levels of moral reasoning because stages five and six were comprised of stereotypically male characteristics. Had Kohlberg incorporated empathy into his stages, maybe women could have better demonstrated their moral reasoning development.

Gilligan challenged previous definitions of morality, many of which only concerned individual justice, rights, and liberties. When studying women's discussions about having an abortion, Gilligan noted that women generally focused on the needs of *others*, rather than their own personal rights or liberties (Tong, 1998). She proposed a novel way to think about the meaning of morality. Through her "morality of care,"

Gilligan (1982) shifted the focus away from justice, rights and liberties to interconnectedness, responsibility for others, and pacifism.

Gilligan's theories of empathy and moral reasoning development in women reformed the conceptualization of normative ethics. Consequently, ER was defined by joining Gilligan's theories about the role of empathy in moral reasoning with other theories discussed in this review. When handling ethical dilemmas concerning empathy, Gilligan would encourage individuals to consider "*How would I respond if I cared deeply about those involved?"* Her ideas about empathy provided a more inclusive, comprehensive definition of ER.

Character. Lastly, consider how Aristotle's philosophies supported to the definition of ER. The word "character" originates from the Greek "charaktêr," and it refers to the distinctive characteristics that set individuals apart from each other (Homiak, 2011). Aristotle's *Nicomachean ethics*, described moral and intellectual virtues of character (Cahn & Markie, 2002; Homiak, 2011). Aristotle explains that intellectual virtues can be obtained, over time, through teaching and training, whereas moral virtues require practice and habit. In addition to practice and habit, people of virtuous character must love virtuous acts, "…just acts are pleasant to the lover of justice and in general virtuous acts to the lover of virtue…; and various actions are such, so that these are pleasant for such men as well as in their own nature" (*Nicomachean ethics*, II. 7). Therefore, a person thought to have a virtuous, or ethical character should love and *choose* to perform virtuous acts. Moreover, the ethical person should find such virtuous acts pleasant.

The definition of ER incorporated Aristotle's theory of the "virtuous" or ideal self. His concepts can be used to confront ethical dilemmas that incite questions about an individual's character. Practicing ER, according to Aristotle, requires asking the question, **"***What actions will help me become my ideal self*?" The concept of one's ideal self is most closely related to Aristotle's "virtuous character."

From this review of various philosophical perspectives, it is clear that ER can be conceptualized in various ways. The Madison Collaborative incorporated each of these philosophical perspectives to support its definition of ER as a robust process. Rather than relying on one theory or philosophy, the Madison Collaborative established ER as a process that emphasizes a set of common ethical considerations. Recall, integrating multiple perspectives in this way represents a casuistry approach.

Examining Different Measures of ER

The aforementioned theories represent an array of ER research and literature. It is clear that researchers defined and measured ER in diverse ways. To understand how previous research conceptualized and assessed ER, consider the following ER measures: The Moral Judgment Interview, the Ethical Reasoning Inventory, the Defining Issues Test, and the Multidimensional Ethics Scale.

Lawrence Kohlberg's moral judgment interview. Kohlberg used an interview process to study stages of moral reasoning development in children and adolescents. Each Moral Judgment Interview (MJI) consisted of three moral dilemmas and standardized questions that asked participants to explain what one *should* do to address the dilemma (Colby & Kohlberg, 1987; Kohlberg, 1958). The three dilemmas in the MJI dealt with life versus law, morality versus punishment, and authority versus upholding an

agreement. Each dilemma had predetermined "central issues." For instance, the MJI used the classic "Heinz" dilemma, and the prescribed central issues were preserving life versus abiding by the law (Colby et al.,1983). Thus, the standardized questions for this dilemma were written to evoke participants' notions of the relative values of human life versus abiding by the law. Kohlberg's MJI initially used Global Story and Sentence Rating scoring systems (Colby et al.,1983). However, inconsistencies in these scoring procedures warranted development of a new scoring method: Standard Issue Scoring (Colby et al., 1983). During the interview process, how a participant responded to the ethical dilemmas allegedly indicated their present stage of moral reasoning. The *correct* answer to Kohlberg's moral dilemma was always the answer that reflected stage six, the highest stage of moral reasoning development (Kohlberg, 1969).

Some reliability estimates for the MJI were favorable; however, several studies did not offer specific information about how reliability estimates were calculated. Therefore, these reliability estimates should be interpreted with caution. Internal consistency reliability estimates for the MJI ranged from .92 to .96 (Colby et al., 1983). Reported test-retest reliability coefficients for the MJI were moderately high, ranging from .96 to .99, with the interval between testing times ranging from three to six weeks (Colby et al., 1983). The MJI has three different forms; five raters independently rated 20 interviews based on one of these forms (i.e., form A). The correlation between raters 1 and 2 on form A of the MJI, or the interrater reliability, was .98, with complete agreement, based on a thirteen-point scale, ranging from 53% to 63%. Additionally, interrater reliability between two raters on forms B and C of the MJI was .96 and .92, respectively (Colby et al., 1983). Given there were three forms of the MJI, alternate form

reliability was also reported by Colby and colleagues (1983); sixty-seven percent of the participants received identical interview scores on two forms of the MJI. Interview scores from Forms A and C of the MJI were correlated .82 and scores from Forms B and C were correlated .84 (Colby et al., 1983). Colby and colleagues (1983) reported "almost all interviews were scored within one-third stage of each other by any two raters, and on about one-half to two-thirds of the interviews the two raters assigned identical scores…" (p. 21).

Critics of Kohlberg's MJI suggested the measure was too difficult to score and administer (Elm & Weber, 1994). The complicated Standard Issues Scoring procedure used to score the MJI required subdividing responses into multiple categories then matching them to detailed criteria from a scoring manual (Colby et al., 1983). Similarly, Maitland and Goldman (1974) thought the MJI had "cumbersome procedural requirements" (p. 700). Although studies suggested the MJI had high internal consistency, the measure was further faulted for lacking acceptable internal consistency and being too subjective (Page & Bode, 1980). Another shortcoming of the MJI was its administration. It required a trained interviewer and it could not be administered to multiple subjects at once because it consisted of a one-on-one interview process (Elm & Weber, 1994). Qualms with the MJI prompted Page and Bode (1980) to develop the Ethical Reasoning Inventory (ERI).

Ethical Reasoning Inventory. The Ethical Reasoning Inventory (ERI) was created to measure moral reasoning in the *exact* same way as Kohlberg's MJI did without the subjectiveness, difficult scoring system, and administration inefficiencies characteristic of the MJI (Page & Bode, 1980). The full version of the ERI contained six

ethical dilemmas. Akin to the MJI, each dilemma used Kohlberg's standard probe questions (Colby & Kohlberg, 1987; Kohlberg, 1958). Page and Bode created 26 questions based on the MJI standard scoring manual. Then they created five written responses to each question that aligned with Kohlberg's first five stages of moral development. The ERI used a "branching technique" to capture participants' responses (Page & Bode, 1980, p. 321). Participants provided their initial "yes" or "no" response to the dilemma then they were redirected to another page of the test depending on their initial answer. If they answered "yes" initially, they were redirected to five different responses or rationales to explain why they selected "yes." If they answered "no" initially, they were redirected to a different set of five responses or rationales to explain why they selected "no." The ERI was scored by calculating the average stage of moral reasoning a participant selected. For instance, stage 1 responses were scored a 1; stage 2 responses were scored a 2, etc. Therefore, a participant that more frequently endorsed higher stage response options would receive a higher score on the ERI.

According to Page and Bode (1980), the ERI showed "high item level consistency" (p. 325). For the full version of the ERI, Cronbach's alpha was .69. After removing the least related dilemma, Cronbach's alpha increased to .74 (Page & Bode, 1980). They also reported a test-retest reliability of .69 for a sample of college students administered the ERI (Page & Bode, 1980). Interestingly, scores on the ERI were correlated .54 with scores on the MJI (Page & Bode, 1980). Recall, the impetus for creating the ERI was the need for an instrument that measured moral reasoning in the same way as the MJI, but with less subjectiveness, easier scoring, and more feasible administration (Page & Bode, 1980). The moderate correlation between ERI and MJI

scores suggested that the ERI might be measuring something that is related to, but not the exact same as, what the MJI measures. Even being mindful of potential method effects, one would expect a higher correlation than .5 if these instruments did indeed tap into the same construct.

The Defining Issues Test. In 1974, James Rest and colleagues from the Center for the Study of Ethical Development at the University of Minnesota created the Defining Issues Test (DIT) to study moral reasoning development (Rest et al., 1974). Kohlberg's method of presenting subjects with a moral dilemma and using their responses to indicate their stage of moral reasoning served as a prototype for creating the DIT. Similar to Kohlberg's moral interviews, the DIT used six different ethical dilemmas to evaluate individuals' current stage of moral development (Rest et al., 1974). Each ethical dilemma was followed by a list of statements or considerations, some of which appealed to an individual's preferred schema. Individuals ranked their preferred considerations as more important than others, while statements that seemed irrelevant were ranked lower. According to Rest and colleagues (1974), highly ranked statements represented the kinds of schema that guided individuals' reasoning about ethical dilemmas.

Kohlberg and Rest defined moral reasoning in different ways. Unlike Kohlberg, Rest and colleagues conceptualized moral reasoning as schemas rather than stages, "overlapping waves" rather than rigid steps in a "staircase" (Siegler, 1997). In other words, Kohlberg thought each stage of moral reasoning was separate and distinct; he drew a stringent or "bright line" between each stage. Yet, Rest and colleagues thought each stage could be demonstrated by a range of responses, all of which are different ways to demonstrate the same stage of reasoning (Elm & Weber, 1994).

Although the DIT was based on Kohlberg's stages of moral development, it differed substantially from the MJI. The DIT was designed to be a recognition test, whereas Kohlberg's MJI was a production test (Rest et al., 1974). That is, the MJI required participants to explain what *ought* to be done to best resolve each dilemma. The DIT, on the other hand, gives participants a list of possible responses to the dilemma; they don't have to formulate their own responses (Elm & Weber, 1994). Structuring the DIT as a recognition test, rather than a production test, alleviated the subjective and arduous scoring procedures used in Kohlberg's interview process (Rest, Narvaez, Bebeau, & Thoma, 1999). Engineering the DIT to have more feasible scoring helped it become the most widely used test of moral reasoning skills (Bailey, 2011). According to Bailey (2011), about 500 researchers, on average, used the DIT each year for the past 15 years.

The scoring system for the DIT focused on Kohlberg's "principled" stages of moral reasoning. The DIT was usually evaluated based on P scores, which represented "the percentage of the respondent's chosen arguments that represent a principled level of reasoning" (Bailey, 2011; Elm & Weber, 1994). According to this rating system, reasoning at any stage other than Kohlberg's principled level was "incorrect." Given that P scores only give credit to statements representing the principled level of reasoning, a student consistently reasoning at any other stages would receive a negligible P score. This scoring method reinforced the hierarchical nature of Kohlberg's stages, maintaining stages five and six as the exemplars of moral reasoning.

In terms of the reliability of DIT P scores, Davison and Robbins (1978) found that P scores had good internal consistency reliability estimates, $\alpha = .77$. Moreover, test-retest

reliabilities for P scores on the full-length DIT ranged from .71 to .82 (Davison & Robbins, 1978). Also, Davison and Robbins (1978) reported a value of .76 for Cronbach's alpha for P scores on a shortened version of the DIT that contained only three dilemmas instead of six. Test-retest reliability estimates for P scores on the shortened version of the DIT ranged from .58 to .77 (Davison & Robbins, 1978).

Longitudinal data raised issues about the validity of DIT P scores. Davison & Robbins (1978) suggested that P scores reflected only upper level stages of moral reasoning; therefore, they could not be used to examine development in moral reasoning. This caused concern because P scores were not adequate measures of change, especially in preconventional and conventional levels of moral reasoning. In essence, P scores were found to be "insensitive to longitudinal trends" (Davison & Robbins, 1978, p. 400). Emler and colleagues raised further concerns about the validity of DIT scores. They found that subjects could change their DIT scores by adopting a particular political viewpoint and answering the DIT from that perspective (Emler, Renwick, & Malone, 1983). However, Bailey (2011) found that scores on the DIT were *not* significantly related to accuracy of evaluating another person's political affiliation. Furthermore, participants scoring higher on the DIT were significantly better at ranking the ethical development of other people (Bailey, 2011). Ultimately, the literature contained conflicting conclusions about the validity of DIT P scores.

The Multidimensional Ethics Scale. After identifying shortcomings in unidimensional ER scales, Reidenbach and Robin (1988) developed the Multidimensional Ethics Scale (MES) to measure ER perspectives in business marketing contexts. They identified two shortcomings of ER measures: "These two problems have

to do with the pluralistic nature of moral philosophy and the single global measures which marketers tend to use in obtaining evaluations of marketing activities" (Reidenbach & Robin, 1988, p. 873). Reidenbach and Robin were frustrated by the limitations of global measures of ER because they wanted to know the specific ethical considerations individuals used to make decisions. An individual could use different perspectives when grappling with ethical issues, yet many instruments provided only one global measure of ethics. They reasoned that a complex construct, such as ER, required a broader and more complex measurement tool than a simpler construct would. This led Reidenbach and Robin (1990) to conclude, "…a multidimensional and multi-item measure seems to be needed to adequately represent this latent construct" (p. 640).

To develop a multidimensional measure of ER, Reidenbach and Robin (1988) presented undergraduate marketing students with three different ethical scenarios and asked them to rate the action of an individual in the scenario in relation to 29 different "philosophy" scales. The 29 philosophy scales reflected different philosophical concepts or dimensions including justice, relativist, egoism, utilitarian, and deontology (Reidenbach & Robin, 1988, p. 874). The following is one of three ethical scenarios and action statements Reidenbach and Robin (1988) used:

> *Scenario C: A retail grocery chain operates several stores throughout the local area including one in the city's ghetto area. Independent studies have shown that prices do tend to be higher and there is less of a selection of products in this particular store than in the other locations.*
>
> *Action: On the day welfare checks are received in the area of the city, the retailer increases prices on all of his merchandise*. (p. 874).

Students rated three action statements, similar to the one shown above, from the perspective of 29 different philosophies using a Likert scale that ranged from 1-7. For example, each student rated the action statement above on a scale of 1 (fair) to 7 (unfair), on a scale of 1 (just) to 7 (unjust), on a scale of 1 (efficient) to 7 (inefficient), on a scale of 1 (culturally acceptable) to 7 (culturally unacceptable), and so on until they rated the action statement on a scale of 1 to 7 for all 29 philosophies. Thus, students responded to 29 items for each of the three different action statements. Students also reported the degree to which they felt they would have made the same decision as the person in the scenario.

According to Reidenbach and Robin (1988), each action statement represented a different measure; therefore, three coefficient alphas were calculated (p. 875). They reported that coefficient alpha for all three action statements was acceptable, ranging from .85 to .87 (Reidenbach & Robin, 1988). Reidenbach and Robin attempted to demonstrate validity evidence for the MES by correlating items from the 29 different philosophy scales with the philosophical areas they were supposed to measure. To clarify, first consider only the relativism scales. Of the 29 different philosophy scales, five were relativism scales. Reidenbach and Robin investigated the extent to which all the items from the five relativism scales correlated with each other. They claimed that "the extent to which they converge, operationalized by high intraclass correlations, the items can be said to measure a common ethical philosophy" (Reidenbach & Robin, 1988, p. 875). Stated another way, Reidenbach and Robin thought that stronger correlations between all of the items on the different relativism scales suggested that the items from the different relativism scales measure one common ethical philosophy. The average correlation

between items purported to measure the same philosophical content were weak to moderate: relativism, r = .54, justice, r = .53, utilitarian, r = .42, deontology, r = .31, and egoist, r = .20 (Reidenbach & Robin, 1988). Reidenbach and Robin interpreted this as convergent validity evidence for the various ethical philosophy scales because the items that were supposed to measure relativism, for instance, were on average, moderately correlated with each other. Convergent validity evidence was important because Reidenbach and Robin hypothesized that individuals would use only one ethical philosophy to evaluate an ethical scenario. To test this hypothesis, items purported to measure the same philosophy needed to be strongly correlated with each other; relativism items needed to represent relativism, not other philosophies such as justice or utilitarianism.

Reidenbach and Robin's supposed convergent validity evidence was questionable. Furthermore, they failed to provide evidence of divergent validity because some items that purported to measure one ethical philosophy were highly correlated with items that allegedly measured a different philosophy. They blamed the lack of divergent validity on "possible conceptual overlap of the different moral philosophies" (Reidenbach & Robin, 1988, p. 875).

Reidenbach and Robin (1988) used exploratory factor analysis (EFA) to explore the dimensionality of the MES; factors were not allowed to correlate in order to "provide as maximally different structures as possible" (p. 875). Based on the EFA results, Reidenbach and Robin rejected the hypothesis that individuals used only one philosophical perspective to evaluate an ethical scenario. When evaluating the scenarios, participants did not use distinctive types of philosophies, which Reidenbach and Robin

interpreted as supporting Rest's stages of moral development. Interestingly, participants used similar criteria when evaluating the ethical content of the scenarios and when evaluating what their behavior would have been if they were personally involved in the scenarios.

Overall their findings promote models of ER that do not rely on only one philosophical perspective (Reidenbach & Robin, 1988). That is, the EFA results provided initial evidence to support a multidimensional factor structure. For instance, five factors had eigenvalues greater than one for the action statement associated with Scenario C. Furthermore, Reidenbach and Robin reported factor patterns that emerged for Scenarios B and C did not align with their hypothesized patterns. That is, participants were not using a singular philosophy to evaluate the various ethical scenarios (Reidenbach & Robin, 1988). Reidenbach and Robin (1988) concluded that, "…there appears to be no single standard of evaluation. That is, the nature and organization of ethical evaluative criteria appear to be situation specific." (p. 879).

Reidenbach and Robin spent the next two years piloting the MES, refining the items, and studying the structure of the scale. They administered the MES to business students and retail managers to address item ambiguity, reduce confusing wording, and identify problems related to the three scenarios (Reidenbach & Robin, 1990). Based on these results, the MES was reduced from 29 items to 8 items. Reidenbach and Robin (1990) used principle components factor analysis with varimax rotation to define three dimensions of the MES: moral equity (four items), relativism (two items), and a social contract construct (two items). Reidenbach and Robin (1990) reported coefficient alpha reliability estimates for all three action statements that ranged from .71 to .92; however, it

is unclear whether these reliability estimates were for the 29-item or 8-item version of the MES. Furthermore, if the MES were multidimensional, coefficient alpha would not be the most appropriate reliability estimate (Cronbach, 1951).

Reidenbach and Robin (1990) concluded that individuals used a broad sense of moral equity focused on justice and fairness when evaluating scenarios instead of using a predefined set of utilitarian or egoist principles. Second, a multidimensional measure of ER provides more information than global measures (Reidenbach and Robin, 1990). A respondent would receive a separate score for each of the three dimensions (i.e., moral equity, relativism, and social contract). Finally, respondents' scores on the three different dimensions offered insights about why they judged a given scenario to be unethical (Reidenbach and Robin, 1990). Reidenbach and Robin (1990) claimed that the MES provided specific information in the form of three "dimension" scores that a unidimensional measure could not provide with one overall score.

McMahon and Harvey studied the dimensionality of the 8-item and 30-item MES and found some concerning results. Specifically, they were skeptical that the MES could measure ethical judgments across a range of different scenarios (McMahon & Harvey, 2007). McMahon and Harvey also described issues with the factor analysis technique used to develop the scale. For example, several studies of the MES used principle components factor analysis, which McMahon and Harvey (2007) stated, "assumes that no measurement errors or other construct-irrelevant sources of variance exist" (p. 29). Additionally, results from exploratory factor analyses (EFA) and confirmatory factor analyses (CFA) suggested the MES might not be a "multidimensional" measure. That is, McMahon and Harvey (2007) found supporting evidence for a 1- and 3-factor structure

for the 8-item version of the MES. McMahon and Harvey (2007) also found that the 30-item MES was "dominated by a general factor representing ethical perceptions" (p. 32). If a one-factor model adequately reproduces the observed item covariances, then the MES might not be multidimensional after all.

Synthesis of ER Measures

The previously described measures defined ER in various ways, but none align with the Madison Collaborative's model of the ER process or the eight KQs. The MJI, the ERI, and the DIT are rooted in Kohlberg's conceptualization and definition of moral reasoning. Recall, however, the Madison Collaborative theoretically and empirically defined ER using a casuistry approach that draws from different philosophical viewpoints and does not involve stages. These instruments are not appropriate for use at JMU for several reasons. First, the MJI and ERI are inappropriate measures to assess the Madison Collaborative's SLOs because both defined ER using a singular, stage or schema perspective put forth by Lawrence Kohlberg and James Rest. Moreover, even if the theoretical underpinnings of Kohlberg's MJI better aligned with the Madison Collaborative's definition of ER, the MJI would still be inappropriate for use at JMU due to administration and scoring issues. The administration and scoring procedures for the MJI are far too arduous; it would be unreasonable to administer and score the MJI for a large cohort of JMU students. Additionally, the Madison Collaborative must be able to evidence the effectiveness of several campus-wide interventions. To do so requires an instrument that is sensitive to changes in students' ER abilities over time. Unfortunately, the DIT might not be sensitive enough to capture ER growth over time (Davison & Robbins, 1978).

Second, the MES cannot assess the Madison Collaborative's SLOs because it does not define ER in terms of the eight KQs. That is, the MES defined ER as a multifaceted construct that required a multidimensional measure. However, the Madison Collaborative conceptualized ER as a multi-step process, where each individual step or set of steps in the process (i.e., the SLOs) is distinct and can be measured using a unidimensional instrument.

Lastly, the intended uses of the MJI, ERI, and MES do not align with the Madison Collaborative's assessment needs. Given the MJI and ERI were created mainly for use with children and adolescents, the validity of these test scores when administered to college students is unknown. The MES was developed for use in specific business contexts; therefore, it is not appropriate to measure ER across all three Madison Collaborative application areas (i.e., personal, professional, and civic) (The Madison Collaborative: Ethical Reasoning in Action, 2013, p.66).

The reviewed instruments do no define ER using the Madison Collaborative's framework; are not easily administered or scored in large-scale testing situations; are not sensitive to change over time; and were not developed for the Madison Collaborative's intended uses. Therefore, none of these instruments can be used to assess students' progression through the Madison Collaborative's SLOs (The Madison Collaborative: Ethical Reasoning in Action, p.23, 2013).

The Ethical Reasoning Identification Test.

Upon reviewing the literature, it was apparent that the Madison Collaborative could not use an existing instrument. Therefore, the Madison Collaborative created a new instrument to measure ER that aligned with their definition of the construct. The ERIT is

one of several assessment tools created to this end; others address different SLOs (See Figure 2). The ERIT is a multiple-choice test that requires students to correctly match the key question most consistent with the rationale used to make a decision in an ethical situation. Each item on the ERIT has eight response options, one for each of the eight KQs, and each item maps onto one of three application areas: *Personal*, *Professional*, and *Civic* (The Madison Collaborative: Ethical Reasoning in Action, 2013, p.66).

History of test development. The current version of the ERIT was created through item writing sessions, pilot testing, and evidence-based revisions. Before the item writing process commenced, the Madison Collaborative clearly stated the purpose and intended use of the ERIT (The Madison Collaborative: Ethical Reasoning in Action, p.66, 2013) (Standard 3.2, AERA, 1999). According to the Madison Collaborative,

> This selected-response, or multiple-choice, test is a direct measure aligned with SLO 2 and SLO 3 (selecting KQs). The ERIT is designed to assess students' ability to differentiate and choose among the 8KQs when confronted with an ethical decision or dilemma (addressing SLO 2). (p. 66)

Students cannot fully engage in the process of ER, represented by SLO 5, until they demonstrate the skills described in the lower level SLOs. To reiterate, students cannot achieve SLO 5 unless they know all eight KQs (i.e., SLO 1); can identify the KQs most applicable to a given scenario (i.e., SLO 2); can identify the appropriate considerations for each KQ (i.e., SLO 3); and can weigh and balance all eight KQs (i.e., SLO 4). Therefore, the ERIT does not measure ER as an entire process; it *measures one particular aspect of the ER process*. Given the theory underlying the hierarchical nature

of the SLOs, students that perform poorly on the ERIT will struggle to achieve subsequent, higher level SLOs.

In accordance with Standard 3.5 (AERA, 1999), to generate items for the ERIT five faculty members from different academic disciplines participated in a two-day workshop facilitated by an ethical reasoning expert (The Madison Collaborative: Ethical Reasoning in Action, p.66, 2013). During the workshop, they engaged in item-writing sessions for two to three hours each day. Faculty wrote a total of 125 items that consisted of ethical scenarios and rationale related to each of the KQs. (The Madison Collaborative: Ethical Reasoning in Action, p.66, 2013). After a comprehensive review process, the Madison Collaborative retained 96 of the 125 items (The Madison Collaborative: Ethical Reasoning in Action, p.66, 2013). During the item review process, one faculty member expressed concerns that the ERIT was "merely a vocabulary test." (David McGraw, personal communication, October 23, 2012). He thought that the items evaluated verbal proficiency instead of lower-level ER skills.

Next, assessment consultants from the Center for Assessment and Research Studies (CARS) piloted the 96 items to make data-based decisions about which of the items should be retained (Standards 3.8 & 3.25, AERA, 1999). In fall 2012, assessment consultants administered two pilot versions of the ERIT (TERA and TERB), each containing 48 non-overlapping or non-common items. Data were collected from 918 freshmen students: 469 freshmen were administered the TERA and the remaining 449 completed the TERB. There were 23 cases that contained missing data for the TERA and 17 cases containing missing data for the TERB. Listwise deletion was used to remove these cases, resulting in a sample size of 446 for the TERA and 432 for the TERB.

Assessment consultants used these data to analyze the pool of 96 items and select the 60 best items to form the ERIT-0 (See Appendix A).

Assessment consultants selected items that were functioning well psychometrically. Classical Test Theory (CTT) techniques were used to evaluate the items. The item selection process followed Crocker and Algina's guidelines. For example, items were selected based on Crocker and Algina's discrimination criteria and item difficulty (1986). In adherence with Standard 3.11(AERA, 1999), assessment consultants also selected the best items based on the extent to which they represented the content domain (See Figure 3). To ensure the ERIT-0 covered the breadth of the ER construct, consultants paid careful attention to content coverage, attempting to equitably balance the cells of the content map (See Figure 3). Some items from the TERA and the TERB had adequate psychometric properties; however, they were not selected for the ERIT-0 because there were already enough well-functioning items for the corresponding KQ and application area (Crocker & Algina, 1986). That is, assessment consultants had to consider the psychometric functioning of each item, while also equitably balancing the cells of the content map.

Appendix A provides item analysis information and a rationale as to why certain items performed poorly for the TERA and TERB. In general, some items performed poorly because students confused the KQs. For instance, students were confused about Liberty and Responsibilities in the following item from the TERA:

> *Danny decides not to take part in the U.S. flag burning demonstration. He tells his friends, "I pledged allegiance to this country while looking at the flag. To burn the flag would be to dishonor this pledge."*

Only 20% of students endorsed the correct response for this item. However, the most frequent response was Liberty, followed by Responsibilities. Students seemed to struggle with items corresponding to Liberty and Responsibility. In addition, students often failed to distinguish Liberty from Rights. For example, the following item probably did not perform well because students were confused about the distinctions between Liberty and Rights:

> *When people criticize Sara's ownership of guns, she usually smiles and says, "Please refer to a document called the Constitution and look up the 2nd amendment."*

For this item, students endorsing Liberty, the incorrect answer, scored about as well on the ERIT as students endorsing Rights, the correct answer.

Several items were performing poorly because they were too easy. For instance, 97% of students answered the following item correctly:

> *Anja always made sure that her children received the same amount of presents for Christmas so that none could accuse her of favoritism.*

Students endorsing "self" as the correct answer to this item scored higher, on average, than students providing the correct answer, fairness. For this item and several others with high difficulty values, the correct answer was probably too obvious. Recall, higher difficulty values mean a higher percent of students are endorsing the correct answer. Items that are too easy (i.e. items with large difficulty values) do not discriminate students performing well on the ERIT from low performing students.

Lastly, several items performed poorly simply because they were poorly written items. Each item was written to correspond to one specific Key Question; however, not

all items clearly aligned with their intended Key Question. These items contained confusing wording or unclear language:

1) *Paige grew up practicing Judaism. When she came to college, most of her friends were Christian. She felt it was important that she share her often differing perspective.*
2) *After an especially profitable year, during which his entire team worked really hard, John Doe learned that the boss intended to give him, and only him, a modest year-end bonus. If he pockets the money, he knows he will feel like a schmuck, so he decides to host a party for the entire team instead.*
3) *Although she was close with her coworker, Jessica knew she had to turn him in, once she found a bag of marijuana under his desk.*

Consider the last item about Jessica finding drugs in her co-worker's desk. The item was intended to align with the Authority Key Question; however, the item does not reference an authority figure. Not surprisingly, only 24% of students endorsed the correct answer. Most students indicated that Responsibilities was the correct answer. The item probably aligned more with the Responsibilities Key Question than Authority.

After selecting the 60 best items from the TERA and TERB to form the ERIT-0, assessment consultants administered the ERIT-0 to sophomore students during a mandatory, assessment day at JMU in spring 2013. The standardized procedures for this Assessment day mirrored those used during fall 2012 Assessment day (Standards 5.1 & 5.6, AERA, 1999). Data were collected from 831 sophomore students with 45-70 earned credits. Listwise deletion was used to remove 38 cases that contained at least one missing data point, resulting in a sample size of 793.

Exploratory factor analysis. In addition to item difficulty, discrimination, and reliability, Exploratory Factor Analysis (EFA) was conducted to investigate the factor structure of the ERIT-0. Given data were dichotomously scored, 1 for correct and 0 for incorrect, the items and factors were not linearly related; therefore, Pearson correlation coefficients were not appropriate. Additionally, a continuous variable was assumed to underlie the observed categories (correct and incorrect). That is, students possess ethical reasoning skills that fall along a continuum; they do not either have ethical reasoning skills or completely lack them. The former requires use of tetrachoric correlations. EFA was conducted using Mplus software (Muthén & Muthén, 2012). Data were analyzed via weighted least squares mean and variance adjusted estimator using tetrachoric correlations. Oblique rotation was used, allowing the factors to correlate with each other. Although EFA provided measures of statistical fit and relative fit, these indices were not interpreted for this EFA as they would be for a Confirmatory Factor Analysis (CFA). Given the exploratory nature of test development, model evaluation focused more on theory than on statistical fit indices.

A one-factor model was estimated because the ERIT-0 was created to measure a distinctive, lower-level part of the ER process. Table 1 displays the factor loadings for the one-factor solution. All but one item (#44) had a significant factor loading. Based on the substantive theory and the scree plot (See Figure 4), it appeared that the one-factor model offered the most parsimonious solution. If the ERIT truly has a unidimensional factor structure, a total score should be reported for the entire test.

Recall, the ER process also has three areas of application and eight KQs; therefore, 3- and 8-factor solutions seemed plausible and were also tested. The rationale

for investigating a 3-factor model was related to the three application areas; theoretically, each application area would form its own factor. As seen in Table 2, after oblimin rotation was applied, the majority of the items did not load onto what theoretically should have been their common application area. The first and second factor were moderately correlated .539; however, the first and third factors as well as the second and third factors were not highly correlated (r = .030 and .067, respectively). It does not appear that the 3-factor model represents the three application areas because each application area did not form its own factor. For instance, items from the personal, professional, and civic application areas loaded onto the first, second, and third factors. The 3-factor model did not differentiate the three application areas, as theory would have prescribed.

Given each item is written to align with one KQ, theoretically, each item would load onto the factor that corresponds to its KQ. Therefore, each factor would represent one KQ. Factor loadings for the 8-factor model are presented in Table 3. The 8-factor model did not fit the data well because the majority of the items did not load onto their appropriate KQ factor. For example, after oblimin rotation was applied, items that were written to measure the empathy KQ did *not* hang together more with each other than with items written to measure the other KWs (See Table 3). Factors seven and eight were very weakly correlated with the other six factors (See Table 4); the remaining factors were weakly to moderately correlated with each other.

Based on CTT, EFA, and content mapping, assessment consultants selected the 50 best items from the ERIT-0 to form the ERIT-1. Also, seven items were revised based on suggestions from Dr. William Hawk, an ER professor at JMU. Given the EFA results, the

next logical step was to test the internal factor structure of the ERIT-1 using a more rigorous procedure, Confirmatory Factor Analysis (CFA).

Reliability evidence. Estimating the reliability of ERIT scores was an important part of the test development process. Given unreliability biases nearly every statistic, it is crucial that the ERIT yields highly reliable scores (Ree & Carretta, 2006). The reliability of ERIT scores should indicate the extent to which items on the ERIT are intercorrelated. Specifically, reliability estimates provide an index of ERIT score consistency (Traub & Rowley, 1991).

One measure of internal consistency is coefficient alpha (Cronbach, 1951). Coefficient alpha assumes ERIT items are unidimensional and at least tau-equivalent (Cronbach, 1951). Although the Madison Collaborative wrote the ERIT items to represent a lower-level step on the ER process, to be unidimensional, we do not know yet if the items actually function in a unidimensional manner. Furthermore, there is no evidence that the factor loadings are equal across all items (i.e., evidence of tau-equivalence). The values of coefficient alpha are reported (See Table 5); however, any measurement expert would likely interpret these reliability estimates with caution because the factor structure of the ERIT is still unknown.

Calculating coefficient alpha for ERIT scores required the Kuder–Richardson Formula 20 because the ERIT was dichotomously scored, 1 for correct and 0 for incorrect responses (Kuder & Richardson, 1937). As shown in Table 5, assuming a unidimensional factor structure, coefficient alpha (KR-20) for the ERIT-0 was .872, indicating good internal consistency of test items. Coefficient alpha (KR-20) for the ERIT-1 was .809 for the full, 50-item version of the test, and .787 for the reduced form of the ERIT-1 that did

not contain the eight testlet questions. Both estimates of coefficient alpha indicated good internal consistency of items on the test.

To provide a more accurate estimate of internal consistency that does not require tau equivalent items *or* a linear relationship between items and the latent factor, McDonald's (1999) omega can be calculated using a nonlinear Structural Equation Modeling (SEM) methodology (Green & Yang, 2008). McDonald's (1999) omega still requires knowledge of the ERIT facture structure; therefore, this internal consistency estimate should also be interpreted cautiously.

Overall evaluation of the ERIT through Benson's Framework.

Benson's (1998) construct validation framework is used to describe the strengths and diagnose the weaknesses of the ERIT-1. Benson's three-stage model of construct validation includes: 1) a substantive stage that emphasized clear definition of the theoretical and empirical domains of ER; 2) a structural stage focused on internal consistency of items on the ERIT and the dimensionality of the ERIT; and finally 3) an external stage concerned with the relationship between ER and other constructs, and test performance of groups known to possess ER skills.

Strengths. Many strengths of the ERIT come from the test development processes used to create and revise the test. Trained assessment professionals adhered to the *Standards for Educational and Psychological Testing* (AERA, 1999) throughout the test development process. As described in chapter one, the Madison Collaborative worked with ethical reasoning content experts to clearly articulate the conceptual and empirical definition of ER. Through the SLOs, they transformed the theoretical understanding of ER into tangible outcomes that could be observed and assessed; they

operationally defined ER. Then SLOs were carefully mapped to ER assessment instruments (See Figure 2). Faculty and administrators associated with the Madison Collaborative reviewed these test development processes. In addition, a regional accreditation team from the Southern Association of Colleges and School Commission on Colleges (SACSCOC) reviewed the Madison Collaborative's proposal and assessment plan. The SACS COC team included a nationally recognized ER content expert, Dr. Elaine Englehardt. Dr. Englehardt visited campus and met with various members of the Madison Collaborative, including the assessment liaisons. After a comprehensive review, Dr. Englehardt and the other members of the SACS COC review team had no formal recommendations for the Madison Collaborative. Each of these processes aligned with Benson's (1998) substantive stage of construct validation.

After establishing the theoretical and empirical domains of ER, the Madison Collaborative created the ERIT to directly assess the part of the ER process described in SLO 2 and 3, a lower-level learning step in the ER process (The Madison Collaborative: Ethical Reasoning in Action, 2013, p.66). ERIT items were written during a workshop, led by an ER expert. A pool of 96 items was piloted during fall 2012. Data collected across three semesters were used to make revisions and explore the internal structure of the test (The Madison Collaborative: Ethical Reasoning in Action, 2013, p.66).

Additionally, there is initial evidence that ERIT items have high internal consistency assuming a unidimensional factor structure. Recall that Benson's (1998) structural stage focuses on the pattern and strength of the correlations between items, and the dimensionality of the instrument. The structural stage represents how items on the ERIT vary with each other and with the theoretical domain of ER (Benson, 1998). For

example, items on the ERIT were created to measure a lower-level step in the ER process (i.e., SLO 2 & 3), thus one might expect all items on the ERIT to be strongly correlated with each other.

With respect to internal consistency, the ERIT-0 and ERIT-1 had acceptable reliability estimates, $\alpha = .872$ and $\alpha = .809$, respectively. The 50-item ERIT-1 has 10 fewer items than the ERIT-0 and can be administered in less time. Also, the ERIT-1 was challenging, but not too difficult for a group of entering freshmen students at JMU (M = 69.00%, SD = 13.07%). EFA results provided some evidence that a one-factor model offered a plausible solution. Assuming a unidimensional factor structure, Spearman-Brown prophecy formula reliability estimates (See Table 5) suggested that the increase in reliability for the full 50-item version of ERIT-1 was due to improved internal consistency or quality of retained items, not due to increasing test length by two items (i.e. going from a 48-item to a 50-item test). Given a unidimensional factor structure, the reliability for the 42-item version of the ERIT-1 that did not contain the eight testlet items was .787. Notice that the value of coefficient alpha for the 42-item version of the ERIT-1 is higher than the Spearman-Brown predicted reliability estimates for a 50-item version of the ERIT-1 (See Table 5).

Weaknesses. The Madison Collaborative made considerable progress in relation to Benson's Substantive stage through re-imagining the theoretical and empirical definitions of ER, and creating the ERIT to measure a lower-level step in the ER process. Nonetheless, the Madison Collaborative has only nominally addressed the structural stage through CTT and EFA analyses. They need further evidence to support Benson's structural stage of validation (Standard 1.1, AERA, 1999). In order to examine and

provide support for the structural stage, Structural Equation Modeling (SEM) will be used. Specifically, CFA will be used to test the Madison Collaborative's ER theory. That is, theory suggests that a unidimensional factor structure is plausible for the ERIT. The unidimensional factor structure hypothesis will be compared to three competing hypotheses including a 3-factor model which represents each of the three application areas, an 8-factor model which represents each of the 8 KQs, and a 3*8 hierarchical model which represents complex items associated with one KQ and one application area.

As Nunnally (1978) pointed out, establishing the structural stage provides necessary but not sufficient evidence of construct validation. Benson's (1998) final stage of external validation summarizes Nunnally's requirements for "sufficient" evidence. For the ERIT, Benson's external stage is concerned with the relationship between one step in the ER process, as measured by the ERIT, and other constructs. The external stage also involves test performance of groups known to possess the skills needed to engage in a lower-level step in the ER process. That is, groups known to have the ability to engage in a lower-level step in the ER process (i.e., SLOs 2 and 3) should perform well on the test relative to those without such proficiency. The external stage is an essential part of the construct validation process. In fact, Benson (1998) considered the external stage of construct validation to be the "most crucial" because it involves understanding what is *actually* being measured by the ERIT (p. 14). Collecting evidence for Benson's external stage helps build a nomological network which in turn helps to convey what construct is actually being measured; naming the construct "a lower-level step in the ER process" does not mean the ERIT is actually measuring "a lower-level step in the ER process." A major weakness of the ERIT is that it lacks evidence to support Benson's external stage. Given the Madison

Collaborative has yet to address Benson's final stage, the next logical step is to collect external validity evidence.

To provide evidence for Benson's external stage, the relationship between ERIT and SAT "Critical Reading" scores (formerly verbal proficiency scores) should be examined because a criticism of the ERIT is that it measures language abilities rather than ethical reasoning. To address this validity issue and gather divergent validity evidence, I will examine the strength of the relationship between ERIT scores and SAT verbal proficiency (SAT-CR) scores. I expect that these scores would be weakly to moderately related; however, SAT-CR scores should *not* be able to explain a substantial amount of variability in ERIT scores. The ERIT should measure a construct that is somewhat related to, *but not the same as*, verbal proficiency. Examining this relationship will demonstrate whether these two constructs relate in the way theory suggests they should relate. Theory also suggests that the relationship between ERIT scores and SAT-CR scores should be weaker than the relationship between ERIT scores and scores on the Ethical Reasoning Recall Test (ERRT). Both the ERIT and the ERRT were created to measure lower level ER skills whereas SAT-CR scores were intended to measure verbal proficiency.

Regarding convergent validity evidence, ERIT scores should be moderately correlated with scores on another Madison Collaborative assessment instrument that also measures lower-level ER process skills. The ERRT measures students' ability to recall from memory and describe all eight KQs, the first and lowest-level SLO (See Figure 2). Given the SLOs are hierarchical in nature, theory dictates that students should master the abilities measured by the ERRT before they can master the abilities assessed by the

ERIT. Thus, students that perform poorly on the ERRT are probably the same students that perform poorly on the ERIT. Examining the relationship between scores on the ERRT and scores on the ERIT will demonstrate whether these scores relate in the way theory suggests they should relate. Also, this will demonstrate whether ERIT scores are more strongly correlated with another ER measure (i.e., the ERRT) or with the construct of verbal proficiency (i.e., the SAT-CR).

Comparing two groups that are expected to have different amounts of experience engaging in the ER process would provide external validity evidence for ERIT scores. That is, if two groups completed the ERIT and the group expected to have some ER process skills outperformed the group expected to have fewer ER process skills then there would be some indication of external validity evidence for ERIT scores. The ERIT should be administered to a group of students that experienced a "dose" of ER intervention prior to completing the test. This group of students should perform better on the ERIT compared to a group of students that did not experience any ER intervention. Freshmen students that completed the ERIT during Assessment day in fall 2013 experienced a small "dose" of ER intervention before completing the test. Given their exposure to ER intervention during Orientation programming, I expect that this group of students will perform better on the ERIT compared to students that received "no dose" of ER intervention. Specifically, the 2013 freshmen cohort should perform better on the ERIT than students who completed the test during fall 2012 when no interventions were in place.

RQ 1- What is the dimensionality of the ERIT-1?

The primary focus of this thesis was examining the factor structure of the ERIT-1 to help JMU stakeholders better understand how to score the test. The factor structure of the ERIT-1 must be tested through more rigorous procedures such as CFA. That is, the EFA performed on the test examined unrestricted models in which every item was allowed to load onto every factor (Kline, 2011). Also, I used oblique rotation which allowed all of the factors to correlate with one another. With EFA, the solution may be rotated as many times as needed to arrive at a parsimonious solution (Kline, 2011). As opposed to EFA, CFA is a more comprehensive tool to evaluate dimensionality. It provides more diverse tests of global fit, allows for comparisons of alternative models, and gives richer item-level fit information (Kline, 2011). It is only through examining competing models that we can gain support for a specific factor structure. Four plausible, theoretical models were tested: a 1-, 3-, 8-, and 3*8-factor model.

Hypothesized one-factor model. The one-factor model (See Figure 5) was plausible because ERIT-1 items were written to align with a lower-level component of the ER process. More specifically, the ERIT-1 was designed to measure SLOs 2 and 3 which represent one skillset comprised of two highly related, lower-level steps in the ER process. All 42 items should share substantively meaningful, common variance with one overarching "Ethical Reasoning Process" factor. For the one-factor model, there were 861 observations in the tetrachoric correlation matrix and 42 estimated parameters resulting in 819 degrees of freedom.

Hypothesized 3-factor model. The 3-factor model (See Figure 6) represents the three areas of application. Theoretically, each Application Area could share substantively

meaningful systematic variance. For instance, items associated with the Personal application area should share common variance that is not shared with the items from the other two application areas. Each factor in the 3-factor model was allowed to correlate with the other factors in the model. For the 3-factor model, there were 861 observations in the tetrachoric correlation matrix and 45 estimated parameters resulting in 816 degrees of freedom.

Hypothesized 8-factor model. The 8-factor model (See Figure 7) represents the eight KQs that define the ER process. In theory, each item could share meaningful systematic variance with the other items that correspond to the same KQ. That is, each item associated with the Empathy KQ should share substantively meaningful, common variance with the other Empathy items that is not shared with the items from the other seven KQs. Each factor in the 8-factor model was allowed to correlate with the other factors in the model. For the 8-factor model, there were 861 observations in the tetrachoric correlation matrix and 70 estimated parameters resulting in 791 degrees of freedom.

Hypothesized 3*8-factor model. The 3*8-factor model (See Figure 8) specified a path from each item to one KQ and one application area. Given each ERIT-1 item should align with one KQ and one application area, the items might be complex. Responses to ERIT-1 items could share substantively meaningful common variance with a KQ, while simultaneously sharing a different kind of meaningful common variance with an application area. If the CFA results provide support for the theory underlying the 3*8-factor model, scoring the ERIT-1 will become more complicated. For the 3*8-factor model, there were 861 observations in the tetrachoric correlation matrix and 115

estimated parameters resulting in 746 degrees of freedom. Note that the 8 KQ latent variables were allowed to freely correlate with each of the other KQ latent variables, and the 3 application area latent variables were allowed to freely correlate with each of the other application area latent variables. However, the 8 KQ latent variables were not allowed to correlate with any of the 3 application area latent variables.

RQ 2- How do ERIT-1 scores relate to SAT verbal proficiency scores?

Given evidence of an interpretable factor structure for the ERIT-1, a secondary focus of this thesis was to gather divergent validity evidence for the ERIT-1. Recall, the ERIT-1 measures one part of the ER process that involves some degree of verbal proficiency. A criticism of the ERIT-1 is that it measures verbal proficiency, rather than part of the ER process. One faculty member suggested that the ERIT-1 is "merely a vocabulary-matching exercise" (David McGraw, personal communication, October 23, 2012). To address this validity issue and gather divergent validity evidence, the strength of the relationship between SAT-Critical Reading (SAT-CR) scores and ERIT-1 scores was examined. Assuming a unidimensional factor structure, it was expected that these scores would be weakly to moderately correlated; however, SAT-CR scores should not share a substantial amount of variability with ERIT-1 scores. That is, the ERIT-1 should measure a construct that is somewhat related to, but not the same as, verbal proficiency. In addition, theory suggests that the relationship between ERIT-1 scores and SAT-CR should be significantly weaker than the relationship between ERIT-1 scores and scores on another Madison Collaborative assessment instrument- the Ethical Reasoning Recall Test (ERRT). Both the ERIT-1 and the ERRT were created to measure lower level ER skills whereas the SAT-CR was intended to measure verbal proficiency. Examining this

relationship will demonstrate whether ERIT-1 and SAT-CR scores relate in the way theory suggests they should relate.

RQ 3- How do ERIT-1 scores relate to ERRT scores? Is this relationship stronger than the relationship between ERIT and SAT verbal proficiency scores?

Given evidence of a unidimensional factor structure, the relationship between ERIT-1 scores and Ethical Reasoning Recall Test (ERRT) scores was examined to gather convergent validity evidence for ERIT-1 scores. The ERRT measures the first and lowest level SLO (See Figure 2); it assesses students' ability to recall from memory and describe all eight KQs. Given the SLOs are hierarchical, theory dictates that students should master the abilities measured by the ERRT before they can master the abilities assessed by the ERIT-1; a student that has not yet mastered SLO 1 should not be able to master SLOs 2 or 3. In other words, students that perform poorly on the ERRT would be expected to perform poorly on the ERIT-1. Therefore, scores on the ERIT-1 should be fairly strongly, positively correlated with scores on the ERRT.

According to theory, the relationship between ERIT-1 and ERRT scores should be statistically significantly stronger than the relationship between ERIT-1 and SAT-CR scores because the ERIT-1, like the ERRT, measures lower-level ER skills, not verbal proficiency. Data from freshman students that 1) provided their SAT-CR scores, 2) completed the ERIT-1 and 3) completed the ERRT were used to compare the correlation between ERIT-1 and ERRT scores to the correlation between ERIT-1 and SAT-CR scores. Fisher's (1921) r-to-z transformation was used to transform the correlations to z-scores. Then the z-scores were used to conduct Steiger's (1980) test of equality of dependent correlations according to the following formula:

$$Z = \frac{\sqrt{(N-3)}(Z_{y1} - Z_{y2})}{\sqrt{(2 - 2\bar{s}_{y1,y2})}}$$

$$\bar{s}_{y1,y2} = \frac{(r_{12})(1-\bar{r}^2-\bar{r}^2) - \frac{1}{2}(\bar{r})(\bar{r})(1-\bar{r}^2-\bar{r}^2-r_{12}^2)}{(1-\bar{r}^2)(1-\bar{r}^2)}, \text{where } \bar{r}^2 \text{ equals } \frac{1}{2}(r_{y1}+r_{y2})$$

RQ 4- Does a group of students that received a "low dose" of ER intervention perform better on ERIT-1 items than a group that received no intervention?

Lastly, to provide external validity evidence for the ERIT-1, I compared two groups' performance on ERIT-1 items. The first group should possess a small amount of ER skill, but the second group should have negligible ER skills, as measured by the ERIT-1. These two groups experienced different levels or "doses" of ER intervention. The group that received a low "dose" should perform better on the ERIT-1 compared to the group that received no intervention because the group that received a low dose of intervention is expected to have minimal ER abilities. For example, incoming freshmen in fall 2013 experienced a low "dose" of ER interventions through a 75-minute Orientation program and the JMU One Book. In contrast, the incoming freshmen cohort that completed the pilot versions of the ERIT-1 in fall 2012 (i.e. the TERA and the TERB) experienced no ER intervention prior to taking the test. The 2013 incoming freshmen cohort that experienced a "low dose" of ER intervention should perform better on ERIT-1 items compared to students that took the test during fall 2012 when no interventions were in place.

Item difficulty scores (p scores) from two different 48-item pilot versions of the ERIT-1 (i.e., the TERA and TERB) administered in fall 2012 were compared to item

difficulty scores from the 42-item ERIT-1 administered in fall 2013. First, I extracted the thirteen items from the TERA and the thirteen items from the TERB (26 total items) that were directly comparable to 42 multiple-choice items on the ERIT-1. Note that all multiple-choice items from the ERIT-1 do not have corresponding items on the two pilot tests because not all 96 multiple-choice items from the pilot tests were selected for inclusion on the ERIT-1 during the test revision process. Moreover, the ERIT-1 included seven new multiple-choice items that were not piloted on the TERA or TERB. After extracting the thirteen common items from each pilot version of the ERIT-1, there were 26 items used for comparison. The item difficulty (p score) for each pilot test item was compared to the item difficulty for its corresponding ERIT-1 item. According to theory, p scores should be higher for freshmen that completed the ERIT-1 in fall 2013 because they received a low dose of ER intervention whereas the freshmen that responded to the TERA or TERB in fall 2012 received no dose of ER intervention at JMU. In other words, the items should be easier for those students who received a dose of ER intervention.

Chapter III: Method

Measures

Ethical reasoning identification test. As described in Chapter two, the assessment consultants worked with Madison Collaborative content experts to create the 50-item multiple-choice Ethical Reasoning Identification Test (ERIT-1). Recall, the ERIT-1was created to measure a lower-level part of the ER process (i.e., SLOs 2 & 3). Items on the ERIT-1 align with one of eight KQs, and one of three application areas: Personal, Professional, and Civic (The Madison Collaborative: Ethical Reasoning in Action, 2013, p.66). Students must identify the appropriate considerations for each KQ and correctly match the most relevant KQ with the given ethical situation; correct and incorrect responses are scored as 1 or 0, respectively.

Note that items 43 through 50 form two testlets comprised of four items each. That is, items 43 through 46 are all based on one scenario and items 47 through 50 are based on a different scenario. Given these items represent testlets, they are not included in the analyses for this thesis; these testlets require further considerations outside the scope of this thesis. Thus, a total of 42 items were used to examine the factor structure of the ERIT-1 and collect validity evidence for ERIT-1 scores.

Ethical reasoning recall test. The Ethical Reasoning Recall Test (ERRT) is a constructed-response instrument that measures the lowest level SLO (See Figure 2). Content experts from the Madison Collaborative created the ERRT to assess students' ability to recall and describe the eight KQs. As the instruments name suggest, students' recall the KQs from memory and provide a written description of each KQ.

The ERRT is scored is two ways. For the literal recall of the eight words associated with KQs (e.g., authority), a SAS program automates the scoring by comparing the students' responses to the actual KQ word (i.e., "authority," "empathy,". If the match is identical or very close (i.e., responsibility for responsibilities) the program will score each correct response as a "1." All other responses are scored a zero. Evaluating the correctness of the KQ explanations is more complicated. Multiple raters evaluate students' explanations relative to the official explanations of each KQ. In addition, the raters use a one-page elaboration sheet, created by an ethical reasoning expert, to help them score each explanation. For the KQ explanation, students receive 1 point for a completely correct answer, half of a point for partial correctness, and 0 points for incorrectness. A student can receive up to 16 total points on the ERRT: eight points for correctly recalling the words associated with the KQs, and another eight points for correctly explaining each KQ.

For this thesis, only the KQ explanation scoring was used because the KQ recall was not as relevant to what the ERIT-1 measures as the KQ explanation. The ERIT-1 presents students with the KQs; thus, students are not required to recall from memory the KQs in order to respond to items on the ERIT-1. Therefore, students' KQ explanation scores are applicable to the specific SLOs associated with the ERIT-1 (i.e., SLOs 2 &3), whereas the KQ recall scores are not (i.e., KQ recall scores are associated with SLO 1). Using only the KQ explanation scoring, the highest score a student could receive on the ERRT was eight points.

SAT critical reading test. The SAT is a standardized test used for college admissions. Students typically complete the SAT during their junior or senior year of

high school. The College Board, supplier of the SAT, asserts that the SAT serves "as a means of leveling the playing field by letting students from all walks of life demonstrate their academic achievement." For the purpose of this research, I am interested in the critical reading (formerly the verbal proficiency) section of the test because it measures verbal proficiency. SAT Critical Reading (SAT-CR) questions "assess students' ability to draw inferences, synthesize information, distinguish between main and supporting ideas and understand vocabulary as it is used in context." According to The College Board website, "The SAT tests are highly consistent with reliability coefficients that are approximately .90." Researchers from the College Board concluded that SAT scores were a good predictor of first year college GPA (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; Patterson & Mattern, 2013). Concerning first-year GPA, SAT scores and high school GPA had similar predictive utility. The College Board also reported that the Critical Reading portion of the SAT correlated 0.48 with first-year college GPA.

Data Collection Procedures

All data analyzed for this thesis were collected from freshmen students during university-wide assessment days at a mid-sized, Southeastern institution during fall 2012 and fall 2013. Data used to answer Research Questions 1, 2, and 3 were collected during fall 2013, whereas data used to answer Research Question 4 were collected during both fall 2012 and 2013 (See Table 6).

Assessment consultants from CARS trained community members and graduate students to proctor assessment day during an hour long workshop. The proctor training workshop included information about the standardized testing procedures, how to motivate students during the testing session, and specific responsibilities of test

administrators. Consistent with Standards 5.1 and 5.6 (AERA, 1999), students completed assessment instruments in a structured, supervised setting. Prior to Assessment day, every freshman student was randomly assigned to a room in an academic building on campus. On Assessment day, different combinations of assessment instruments were administered in different rooms. Thus, students who completed the TERA, TERB, or ERIT-1 represented a randomly selected sample of freshmen students. Students responded to test items using scantron forms (Standard 1.13, AERA, 1999). Although participation in assessment day was mandatory for all incoming freshmen, the tests are low stakes for these students. That is, students' performance on assessment day tests does not affect their GPA or academic standing.

RQ 1-What is the dimensionality of the ERIT-1?

RQ 1 participants. Responses to the ERIT-1 were used to answer the primary research question. Data were collected from 862 freshmen that responded to the ERIT-1 during a mandatory assessment day in fall 2013.After removing the 33 cases that contained missing data, 829 students comprised the sample analyzed for this study. See Table 7 for this sample's demographic information.

Data screening. Of the 50 items on the ERIT-1 only 42 were used for data analyses because the last eight items on the test formed two testlets made up of four items each. Data were screened for missing values and variability within each item. The data contained one duplicate case; it was removed reducing the sample size from 863 to 862. Of the 862 students that completed the ERIT-1, 33 students left at least 1 item blank. More specifically, 24 students left one item blank, six students left two items blank, one student left three items blank, one student left four items blank, and one student left five

items blank. Thus, there were 48 instances of missing data. To clarify, for all 862 cases across all 42 items, there were 48 instances in which a respondent left an item blank. Items 14, 22, and 25 on the ERIT-1 had the most missing data compared to the other items; each item had four missing data points. Listwise deletion was used to remove the 33 students that left one or more items blank, resulting in a sample size of 829.

Item difficulty. Item difficulty values represent the percentage of students that responded correctly to an item. In essence, "difficulty" values can be thought of as an indication of item "easiness;" when an item has a high difficulty value, a large percentage of students responded correctly to that item. Students responded correctly to 68.2% of the 42 multiple-choice items on the ERIT-1 (i.e., the average score was 28.64 out of a possible 42 points). There was substantial variability in item difficulty values for ERIT-1 items (See Table 8). Item difficulty values ranged from .171 for the most difficult item (i39) to .959 for the easiest item (i21). Interestingly, the three most difficult items were all Rights items. Across the eight KQs, Rights and Liberty items were the most difficult. On average, 47.7% of students responded correctly to the six Rights items, and 55.3% responded correctly to the six Liberty items. In contrast, on average, 85.0% of students responded correctly to the Empathy items, 82.6% responded correctly to the Fairness items, and 83.9% responded correctly to the Character items (See Table 8). For 37 out of 42 items, on average, at least 50% of students provided the correct answer. That is, only five items had difficulty values less than .50. This means that at least 50% of students responded correctly to 88% of the items on the ERIT-1 (i.e. 37 out of 42 items). There was variability in student responses for all 42 items.

Item correlations. Given data were ordered and categorical, each item was not linearly related to the latent factor(s); therefore, Pearson correlation coefficients were inappropriate. Compared to tetrachoric correlations, Pearson correlations underestimate the relationships among the ERIT-1 items. That is, although the ERIT-1 is scored dichotomously, it is assumed that there is a continuum of ER ability underlying scores on the ERIT-1. In other words, ER skills are not all or nothing; students do not have ER skills or completely lack ER skills. Thus, the observed categorical scores on the ERIT-1 approximate an underlying continuum of ER skills (Finney & DiStefano, 2013). Tetrachoric correlations represent the estimated theoretical relationships between items on the ERIT-1 that have an underlying continuous distribution (Finney & DiStefano, 2013).

The observed tetrachoric correlations ranged from 0 to .583. Items 20 and 42 were correlated 0, as were items 36 and 42. The largest observed tetrachoric correlation was between items 17 and 35. One pattern that emerged from the observed correlations was that item 42 had very low correlations with the other 41 items on the test. For instance, the highest observed correlation that any given item shared with item 42 was .193. In short, item 42 did not share much variance in common with the other items on the test. Item 42 was scored as correct if students selected the Rights KQ; however, an ER content expert thought that this item could also be aligned with the Liberty KQ. That is, although item 42 was created to align with the Rights KQ, it might be more aligned with the Liberty KQ. Interestingly, 29% of students indicated that Rights was the correct answer to item 42, whereas 36% of students thought Liberty was the correct answer. Perhaps

item 42 did not share much common variance with the other items on the test because it should have been scored as a Liberty item as opposed to a Rights item.

If the ERIT-1 items represented one "Ethical Reasoning Process" factor (see Figure 5) then the correlations among all of the items would be roughly similar, and patterns consistent with a 3- 8- and 3*8-factor model would not be observed. The majority of the 861 correlations in the observed correlation matrix were correlated between .07 and .27. The observed correlations formed a pattern that was consistent with a unidimensional factor structure.

Patterns of observed correlations consistent with the 3 application areas (See Figure 6) were not observed. That is, items from the Personal Application Area were not correlated more highly with each other than with items from the Professional or Civic application areas. The same was true for the Professional items and the Civic items.

Moreover, no patterns of observed correlations were consistent with the 8 KQs (See Figure 7). Items mapped to a particular KQ did not correlate more highly with other items mapped to that same KQ compared to items mapped to other KQs. For example, all of the Fairness items did not correlate more highly with each other than they did with items from other KQs. Item 31, one of five Fairness items, correlated more strongly with several Authority, Empathy, Character, and Liberty items than it did with other Fairness items.

Patterns of observed correlations that suggested each item was complex (See Figure 8) were not observed. If the 42 items were complex, each item would be more highly correlated with other items that share the same KQ *and* application area. For example, items written to represent the Outcomes KQ and the Civic application area

would be more highly correlated with each other than with items mapped to other KQs and application areas. Items 1 and 7 were written to represent the Outcomes KQ and the Civic application area; however, these two items were weakly, *negatively* correlated. Items 19 and 35 corresponded to Empathy and the Professional application area, yet they were only correlated 0.14. Item 35 was more strongly correlated with 35 other items on the test than it was with item 19.

Given the pattern of observed correlations, I expect that the ERIT-1 will have a unidimensional factor structure. The lack of observed correlation patterns consistent with the 3-, 8-, and 3*8-factor models indicated that these models might not fit the data. That is, the observed correlations did not reflect patterns consistent with items loading onto three application area factors, or eight KQ factors. Nor did the observed correlations contain patterns indicating that the ERIT-1 items were complex (i.e., items load onto a KQ and an application area).

Estimation method. The estimation method should provide the most unbiased, efficient, and consistent parameter estimates given the ordered categorical nature of the data. It was inappropriate to analyze the data using normal theory estimators or Pearson correlation coefficients because ERIT-1 scores were not continuous and normally distributed. That is, estimators such as Generalized Least Squares (GLS) and Maximum Likelihood (ML) are not appropriate for ordered categorical data because they assume continuous data that follow a multivariate normal distribution. Analyzing nonnormal ordered categorical data using ML estimation produces biased χ^2 values, standard errors, and parameter estimates (Finney & DiStefano, 2013).

Robust Diagonally Weighted Least Squares (rDWLS) was used to estimate the hypothesized models because it accounts for the ordered categorical nature of the data by analyzing tetrachoric correlations instead of Pearson correlations or covariance matrices. Like WLS, rDWLS uses the asymptotic covariance matrix of the tetrachoric correlations being analyzed as the weight matrix; however, it does so without having to invert the full asymptotic covariance matrix (Finney & DiSetefano, 2013). More specifically, rDWLS estimation requires inverting only the diagonal of the asymptotic covariance matrix (Finney & DiSetefano, 2013). When sample sizes are small and models are complex, rDWLS outperforms WLS because rDWLS does not have to invert the full asymptotic covariance matrix. Given only a portion of the asymptotic covariance matrix is used to produce unbiased and consistent parameter estimates, these estimates are not efficient (i.e., χ^2 test statistics and standard errors will be biased). Hence, the DWLS χ^2 test statistic and standard errors must be adjusted for this inefficiency using information from the full asymptotic covariance matrix. That is, the rDWLS estimator employs scaling techniques, similar to those used in the Satorra-Bentler (S-B) scaling procedure (Satorra & Bentler, 1994), to adjust the DWLS χ^2 test statistic and biased standard errors of parameter estimates (i.e., a "mean adjustment" is applied to the unadjusted DWLS χ^2 statistic) (Finney & DiSetefano, 2013). Likewise, fit indices are adjusted by using the adjusted (robust) χ^2 statistic.

Assessing model-data fit. An exact fit index and two global fit indices were estimated to assess model-data fit. The DWLS adjusted χ^2 test statistic is an index of exact fit, meaning it tests how well the model *exactly* fits the data (Weston & Gore, 2006). A significant χ^2 value indicates the observed tetrachoric correlation matrix is

statistically significantly different from the model-implied tetrachoric correlation matrix. It is unlikely that the hypothesized models will exactly fit the data. Therefore, two global approximate fit indices were examined: one absolute fit index and one incremental fit index. This reporting approach followed Hu and Bentler's "2 index presentation strategy" (1998; 1999).

The Root Mean Square Error of Approximation (RMSEA) (Steiger, 1990) was evaluated to assess absolute fit (Yu & Muthén, 2002). RMSEA is sensitive to parsimony; simpler models (models with more degrees of freedom) will have lower RMSEA values. RMSEA indicates the amount of model-data misfit *per every one degree of freedom*. Values of RMSEA range from 0 to 1.0, with values closer to 0 indicating better model-data fit. RMSEA is sensitive to misspecified factor loadings also known as "complex misspecification" (Hu & Bentler, 1998). Yu and Muthén (2002) recommended reporting the robust DWLS RMSEA because it was able to control for inflated Type 1error rates better than the Maximum Likelihood RMSEA when data were categorical.

Bentler's Comparative Fit Index (CFI) was evaluated to assess incremental fit. The CFI compares the fit of the hypothesized models to that of a null model, which specifies no relationships among the 42 ERIT-1 items. The CFI represents the proportion of lack of fit due to the null model. The value of the CFI ranges from 0 to 1.0, with a value 1.0 indicating that the lack of fit is entirely due to the null model; thus, the hypothetical model represents the data well.

Yu and Muthén (2002) found that RMSEA and CFI had more power to detect misspecified factor loadings (i.e., complex misspecifications) than to detect misspecified factor covariances (i.e., simple misspecifications) when data were binary. Yu and Muthén

(2002) recommended cutoffs of .05 and .95 for the RMSEA and CFI, respectively. However, these cutoffs should not be overgeneralized (Nye & Drasgow, 2011). Lower values of RMSEA and higher values of CFI will indicate better approximate fit. Yu and Muthén (2002) did not recommend reporting the Standardized Root Mean Square Residual (SRMR) for dichotomous data because type 1 errors were inflated when sample sizes were less than or equal to 250. Also, SRMR lacked adequate power to reject complex misspecified models when sample sizes were greater than or equal to 500 (Yu & Muthén, 2002). Furthermore, no adequate cutoff value could be established for the SRMR.

Given global fit indices can potentially "mask" areas of local misfit, tetrachoric correlation residuals were estimated to assess discrepancies between observed correlations and model-implied correlations. Higher absolute values indicate greater discrepancy between the observed and model-implied tetrachoric correlations. In other words, higher tetrachoric correlation residual values indicate areas of local misfit or specific relationships that the hypothesized model did not reproduce well. For this study, correlation residuals higher than |.2| indicated areas of local misfit.

A tetrachoric correlation matrix (See Excel file for matrix) was analyzed using LISREL version 8.80 (Jöreskog & Sörbom, 2006). For the one-factor model, the variance of the latent factor was fixed to 1 to set the metric. For the remaining models, the metric of each latent variable was set by fixing the factor pattern coefficient of one indicator to 1. When this methodology was used to set the metric of the latent variables, LISREL provided the standardized solution, which rescaled the factor pattern coefficients as if the

latent variables had a variance of 1. Thus, we get more information for the 3-, 8- and 3*8-facotr models using this method to set the variance of the latent factors.

RQ 2- How do ERIT-1 scores relate to SAT verbal proficiency scores?

RQ 2 participants. SAT Critical Reading (SAT-CR) data were used to answer research question two (i.e., How do ERIT scores relate to SAT verbal proficiency scores?). As part of the college application process, students authorized the College Board to release their SAT-CR (formerly verbal proficiency) scores to the Office of Admissions. Of the 829 students tested during fall 2013 that had complete data on the ERIT-1, JMU had SAT-CR records for 772 of them. Therefore, these 772 students' responses were used to compute the correlation between ERIT-1 and SAT-CR scores.

Assuming evidence supports a unidimensional factor structure, to address Research Question 2, I examined the Pearson product moment correlation between ERIT-1 scores from fall 2013 and SAT-CR scores. Theory suggests that ERIT-1 scores should only be weakly to moderately correlated with SAT-CR scores. The relationship between ERIT-1 scores and SAT-CR scores should not be as strong as the relationship between ERIT and ERRT scores. That is, the ERIT and the ERRT should be measuring ER, whereas the items on the SAT-CR should be measuring a different construct.

RQ 3- How do ERIT-1 scores relate to ERRT scores? Is this relationship stronger than the relationship between ERIT and SAT verbal proficiency scores?

RQ 3 participants. Data used to analyze the ERRT were collected from 255 incoming freshmen students on a university-wide Assessment day in fall 2013. Of the 255 students administered the ERRT, 140 also completed the ERIT-1 on the same Assessment day during fall 2013. Data from the 140 students that completed both

instruments will be used to address research question three (i.e., How do ERIT-1 scores relate to ERRT scores? Is this relationship stronger than the relationship between ERIT and SAT verbal proficiency scores?).

Given evidence supporting a unidimensional factor structure, the third research question was answered by examining the Pearson Product Moment correlation between ERIT-1 scores and ERRT KQ explanation scores. The relationship between ERIT-1 and ERRT scores should be significantly stronger than the relationship between ERIT-1 scores and SAT-CR scores. Data from 140 freshman examinees that provided their SAT-CR scores, completed the ERIT-1, and completed the ERRT were used to compare these two correlation coefficients. Fisher's (1921) r-to-z transformation was used to transform the correlations to z-scores. Then the z-scores were used to conduct Steiger's (1980) test of equality of dependent correlations according to the following formula:

$$Z = \frac{\sqrt{(N-3)}(Z_{y1} - Z_{y2})}{\sqrt{(2 - 2\bar{s}_{y1,y2})}}$$

$$\bar{s}_{y1,y2} = \frac{(r_{12})(1-\bar{r}^2-\bar{r}^2) - \frac{1}{2}(\bar{r})(\bar{r})(1-\bar{r}^2-\bar{r}^2-r_{12}^2)}{(1-\bar{r}^2)(1-\bar{r}^2)}, \text{where } \bar{r}^2 \text{ equals } \frac{1}{2}(r_{y1}+r_{y2})$$

Theory implies that ERIT-1 scores should be statistically significantly more strongly correlated with ERRT scores than with SAT-CR scores.

RQ 4- Does a group of students that received a "low dose" of ER intervention perform better on ERIT-1 items **than a group that received no dose of intervention?**

RQ 4 participants. Data collected during 2012 and 2013 were used to address Research Question 4 (i.e., Does a group of students that received a "low dose" of ER

intervention perform better on the ERIT than a group that received no intervention?). In fall 2012, two versions of the ERIT-1 were administered, the TERA and the TERB. The TERA and TERB each contained 48 items, and they shared no common items. Thirteen items on the ERIT-1 were also on the TERA. Similarly, the ERIT-1 and the TERB shared 13 common items. Hence, given a unidimensional factor structure, scores fromERIT-1 items administered in 2012 and 2013 were compared.

Data were collected from 918 freshmen students administered the TERA or TERB during fall 2012 and from 829 freshmen administered the ERIT-1 during fall 2013. As one might expect, given random assignment and large sample sizes, the demographic information for students completing the TERA, TERB, and ERIT-1 was very similar (See Table 7).

To address the fourth research question, item difficulty scores (p scores) from the TERA and TERB were compared to item difficulty scores from the ERIT-1. Items that the TERA shared in common with the ERIT-1 were extracted; the same was done for the TERB. In total, 26 items from the TERA and the TERB were directly comparable to corresponding ERIT-1 items. Note that all 50 items from the ERIT-1 do not have corresponding items on the TERA and TERB. Recall, the TERA and TERB were pilot tests, thus not all items from the TERA and TERB were selected for inclusion on the ERIT-1. Nine of the items selected from the TERA and TERB for inclusion on the ERIT-1 were revised to address confusing wording; therefore, they are no longer considered common items. Moreover, the ERIT-1 includes seven new items that were not piloted on the TERA or the TERB. Comparing p scores on an item-by-item basis represents the average change in difficulty for ERIT-1 items compared to TERA and TERB items.

According to theory, p scores for the ERIT-1 items should be higher than p-scores for the TERA and TERB items because the freshmen that responded to the ERIT-1 experienced a low dose of ER intervention whereas the freshmen that responded to the TERA or TERB received no ER intervention at JMU.

Chapter IV: Results

RQ 1- Confirmatory Factor Analysis

Recall that answering RQ 1 (i.e., What is the dimensionality of the ERIT-1?) required estimating several CFA models including a 1-, 3-, 8-, and 3*8-factor model. The one-factor model was the only model that converged to an admissible solution. For the three-factor model, the phi matrix was non positive definite. That is, the correlations between the three factors were inadmissible values (i.e., $r_{personal,professional}$=1.06, $r_{personal,civic}$=1.05, $r_{professional,civic}$=1.03). Given three factors may not exist (hence the inadmissible phi matrix that housed the factor correlations), three additional two-factor models were tested that represented all possible combinations of the three factors to assess if a simpler solution may represent the data. For example, a two-factor model in which the Personal area combined with the Professional area constituted one factor and the Civic area constituted the second factor was tested. Each of the three two-factor models resulted in non positive definite phi matrices where the correlations between the two factors were inadmissible (i.e., $r_{personal,pro+civic}$=1.04, $r_{civic,personal+pro}$=1.02, $r_{professional,personal+civic}$=1.03). Thus, the theory underlying the three application areas was not supported.

The eight-factor and 3*8-factor models did not converge to an admissible solution after 50 iterations. For both models, LISREL indicated that the phi matrix was non positive definite. The results suggested that the 3-, 8-, and 3*8-factor models were empirically underidentified (Rindskopf, 1984). It is plausible that the 3- 8- and 3*8-factor models were empirically underidentified due to overfactoring the data (Rindskopf, 1984). Importantly, empirical underidentification is consistent with the hypothesis that the

ERIT-1 is unidimensional. The additional factors were not identified; hence, the 3-, 8-, and 3*8-factor models were inadmissible because the extra factors likely do not exist.

Although the one factor model did not fit the data in an exact sense (DWLS scaled $\chi^2(819)=1245.12$, $p < .001$) the approximate fit indices indicated adequate global fit. The values of the adjusted RMSEA and adjusted CFI were 0.03 and 0.93, respectively. The RMSEA fell within Yu and Muthén's (2002) recommended cutoff of .05 and the CFI approached their recommended cutoff of .95. Table 9 presents the standardized factor pattern coefficients and variance explained for the one-factor model.

The factor pattern coefficients for all but two items were statistically significant (See Table 9). Items 7 and 42 were the only items with non-significant factor loadings, which was not surprising because the observed correlations between item 7 and the other items were low; the same was true for the observed correlations between item 42 and the other items on the test. Item 7 was written to align with the Outcomes KQ, but according to an ER content expert it references a "legal" issue and thus could be aligned with another KQ. Although 57% of students endorsed the correct response for item 7, 18% of students thought that the Responsibilities KQ was the correct answer. As described previously, item 42 was written to align with the Rights KQ; however, based on student responses and feedback from an ER content expert, item 42 might be more aligned with the Liberty KQ than the Rights KQ. Moreover, item 42 is long and given it is located at the end of the instrument, testing fatigue might have affected students' responses to this item.

The factor pattern coefficients ranged from .05 to .73, and they are interpreted as correlations because in the one-factor model only one factor is directly affecting

responses to each item. For example, the squared factor pattern coefficient represents the amount of variance in an ERIT-1 item that is explained by the latent "ER Process" factor. For item 35, this means that 53% of the variance in this item is explained by the "ER Process" factor. Eighteen of the 42 items had factor pattern coefficients greater than or equal to .5; thus, for these 18 items, at least 25% of their variance was explained by the overarching "ER Process" factor.

Diagnosing local misfit. Although the one-factor model had decent global fit, there were some relationships that the one-factor model did not reproduce well. That is, tetrachoric correlations residuals identified areas of local misfit. The correlation residuals ranged from 0 to .275. Only fourteen of the 861 correlation residuals were greater than |.2|. Based on the residuals, local misfit appeared to be mainly associated with KQs (See Table 10). Of the fourteen residuals that were greater than |.2| for the one-factor model, four shared a common KQ, two shared a common Application Area, and one shared both a common KQ *and* a common Application Area (See Table 10). However, no patterns emerged that suggested the local misfit was mainly due to a specific KQ, which suggests that items aligned with the same KQ do not tend to correlate with each other above and beyond their correlations with items from other KQs. Thus, 8 different KQ factors probably do not exist. Alternatively, if the residuals had been associated with a specific KQ, then the items aligned with that KQ might share substantively meaningful relationships above and beyond their relationship with the overall "ER Process" factor.

Given the residuals were not associated with one KQ or application area, in general, items from the same KQ did *not* want to correlate more highly with each other than with items from different KQs; the same is true of the application areas. That is,

items from the same application area did *not* want to correlate more highly with each other than with items from different application areas. If the residuals had been mainly associated with one particular KQ, say Empathy, this would suggest that the Empathy items share meaningful common variance, above and beyond the variance they share with the overarching "ER Process" factor. Overall, no patterns emerged from examining the correlation residuals that supported the existence of a 3-, 8-, or 3*8-factor structure.

LISREL provided modification indices that suggested ways to improve model-data fit by allowing error variances between certain items to correlate (See Table 11). These suggested post hoc modifications are mathematically rather than theoretically based, thus they should not necessarily be used (MacCallum, Roznowski, & Necowitz, 1992). However, the modification indices, much like the correlation residuals, can provide insights into areas of local misfit. For instance, the modification indices revealed that five of the six Liberty items on the ERIT-1 wanted to have correlated error terms. In other words, five of the six Liberty items appeared to share *something* in common over and above the variance they shared with the other 37 items on the test (i.e., the overarching "ER Process" latent factor).

Note that several of the correlation residuals presented in Table 10 also appear in the modification indices presented in Table 11 due to the fact that the modification indices include the correlation residuals. The most noticeable difference between Tables 10 and 11 is that the modification indices are more numerous than the correlation residuals because modification indices are more sensitive to smaller amounts of model-data misfit. Given limited evidence from the modification indices that the Liberty items might share common variance unrelated to the common variance they shared with the

non-Liberty items on the test, a post hoc bifactor model was tested (See Figure 9). Although the modification indices indicated that five of the six Liberty items wanted to have correlated error terms, the Liberty items did not have large correlation residuals. Thus, conducting a test of the bifactor model was conservative. Note the metrics of the latent factors for the bifactor model were set by fixing the variance of the latent factors to 1, and the correlation between the two latent factors was set to 0. The exact fit of the bifactor model was poor but better than the unidimensional model, DWLS scaled $\chi^2(813)=1075.97$, $p < .001$; the global fit indices suggested adequate absolute and incremental fit ($RMSEA_{adjusted}=.02$, $CFI_{adjusted}=.96$).

Table 12 displays the factor pattern coefficients for the one-factor model compared to the bifactor model. First note that, when the subfactor ("Liberty*") was modeled, the factor pattern coefficients of the non-Liberty items onto the "ER Process" factor did not change drastically. That is, adding in the "Liberty*" subfactor did not dramatically change the relationships between the other items and the "ER Process" factor. This pattern provides evidence that the ERIT-1 could be considered *essentially* unidimensional.

Second, note that the factor pattern coefficients for the six Liberty items onto the "ER Process" factor ranged from 0.34 to 0.52, and the loadings for the Liberty items onto the "Liberty*" factor ranged from 0.27 to 0.64. Of the six Liberty items, 5, 14, and 20 related slightly more strongly with the "Liberty*" factor than the overall "ER Process" factor, whereas items 24, 32, and 33 loaded slightly more strongly with the "ER Process" factor .This pattern provides evidence that the six Liberty items shared non-negligible, common variance that was distinct from the variance they shared with the other items on

the test. Expectedly, the bifactor model reduced this misfit among the Liberty items that was observed in the one-factor model (See Table 13 for correlation residuals).

Interestingly, the correlation residuals for the bifactor model revealed a similar pattern that was observed in the one-factor correlation residuals (See Table 10). That is, the fourteen items that had correlation residuals greater than |.2| in the one-factor model also had correlation residuals greater than |.2| in the bifactor model (See Table 10). The average size of the residuals was similar for the one-factor and bifactor models. This finding is not surprising given the bifactor model did not address the largest correlations residuals from the one-factor model (See Table 10). Moreover, the bifactor model was not expected to fit much better than the one-factor model because the Liberty items that comprised the subfactor did not have large correlation residuals.

Given the global and local fit of the one-factor model, the minor change in the item relationships with the "ER Process" factor when the Liberty* subfactor was included in the model, and the limited items showing complex structure, the items were deemed essentially unidimensional. Clearly, the bifactor model should be interpreted cautiously because the one-factor model was post-hoc modified to improve its fit to this single sample; thus, the resulting bifactor solution may not generalize to other samples (MacCallum et al., 1992). However, even if the results for the bifactor model were replicated using another sample, the common variance that the Liberty items shared above and beyond the common variance they shared with the "ER Process" factor was minor, therefore, supporting the use of a single, total score for the ERIT-1.

Based on CFA results, the one-factor model fit the data well enough to support an *essentially* unidimensional structure; therefore, a reliability estimate for the ERIT-1 was

calculated and further analyses were conducted using a total score for the ERIT-1. To accommodate the dichotomous nature of ERIT-1 scores, McDonald's (1999) omega was calculated using a nonlinear SEM methodology described by Green and Yang (2009). Their formula used the thresholds, factor pattern coefficients, and model reproduced tetrachoric correlations to compute a nonlinear estimate of omega (Green & Yang, 2009). Based on Green and Yang's (2009) formula, omega for the one-factor model was .792. The average proportion of variance in the 42 items accounted for by the "ER Process" factor was 19.86%. Given adequate fit of the one-factor model, ERIT-1 total scores were analyzed to gather validity evidence for ERIT-1 scores.

RQ 2- Divergent Validity Evidence

The average SAT Critical Reading (SAT-CR) score for the 772 freshmen students that completed the ERIT-1 and provided SAT-CR scores was 571 (*SD*=67.43). The average ERIT-1 total score for the 772 freshmen was 28.77 (*SD*=5.69). Total scores on the 42 multiple-choice items on the ERIT-1 were moderately correlated with SAT-CR scores, $r(772)=.439$, $p<.001$. Squaring this correlation indicates the proportion of variance in ERIT-1 scores that is shared with SAT-CR scores. In other words, 19.3% of the variance in ERIT-1 scores is shared with SAT-CR scores. Although the correlation was significant, nearly 81% of the variability in ERIT-1 scores cannot be explained by SAT-CR scores, suggesting that ERIT-1 scores represent something different than SAT-CR scores. Thus, there is evidence that the ERIT-1 measures a construct that is somewhat related to, but not the same as, verbal proficiency.

RQ 3- Convergent Validity Evidence

The average Ethical Reasoning Recall Test (ERRT) score for the 140 freshmen students that responded to both the ERIT-1 and the ERRT was 4.21 out of a possible 8 total points (*SD*=1.80), meaning that on average students could correctly define or explain approximately four of the eight KQs. The average ERIT-1 total score for the 140 freshmen students was 29.81 (*SD*=5.83). Total scores on the 42 multiple-choice items on the ERIT-1 were significantly, weakly correlated with ERRT scores, $r(140)=.257$, $p=.002$. Squaring the correlation indicated that 6.6% of variance in ERIT-1 scores is shared with ERRT scores. The correlation between ERIT-1 and ERRT scores represents only a small effect, $r < .3$ (Cohen, 1992). The fact that the correlation between ERIT-1 and ERRT scores is positive and statistically significant supports the theory that students performing well on the ERIT-1 should be the same students that perform well on the ERRT. However, theory suggests that the correlation between ERIT-1 and ERRT scores should have been substantially stronger than the correlation between ERIT-1 and SAT-CR scores. The ERIT-1 and ERRT should measure lower levels steps in the ER process (i.e. SLOs 1, 2, &3), whereas SAT-CR scores should measure verbal proficiency; thus, ERIT-1 and ERRT scores should be more highly correlated than ERIT-1 and SAT-CR scores.

Steiger's (1980) test of equality of dependent correlations indicated that the correlation between ERIT-1 and SAT-CR scores was statistically significantly stronger than the correlation between ERIT-1 and ERRT scores, $z = -4.39$, $p < .001$. To further test the convergent and divergent validity of ERIT-1 scores, data from 122 students who 1) responded to the ERIT-1, 2) responded to the ERRT, *and* 3) provided SAT-CR scores

were gathered. The correlation between ERIT-1 and ERRT scores was weak $r(122) = .258$, $p = .004$, as was the correlation between ERRT and SAT-CR scores, $r(122) = .288$, $p = .001$, whereas the correlation between ERIT-1 and SAT-CR scores was strong, $r(122) = .651$, $p < .001$ (Cohen, 1992). This finding did not support the theory that ERIT-1 and ERRT scores would be statistically significantly more strongly correlated than ERIT-1 and SAT-CR scores.

From the above information, I further examined the relationship among ERIT-1, ERRT, and SAT-CR scores. Perhaps ERRT scores still had a unique relationship to ERIT-1 scores, which could best be explored by examining the correlation of ERRT to ERIT-1 scores *after controlling for* verbal proficiency (i.e., SAT-CR scores). Using multiple regression, SAT-CR scores were entered into the model first to predict ERIT-1 scores then ERRT scores were subsequently entered into the model to examine their predictive utility *above and beyond* SAT-CR scores. ERRT and SAT-CR scores together accounted for a significant percentage of variance in ERIT-1 scores, $R^2 = .655$, $F(2,121) = 44.769$, $p < .001$, 95% CI: .54 to .74 (Steiger & Fouladi, 1992); however, only SAT-CR scores contributed significantly to the model ($b = .055$, $p < .001$, 95% CI: .042 to .068, $sr^2 = .362$). Thus, ERRT scores did not significantly predict ERIT-1 scores, after controlling for SAT-CR scores ($b = .259$, $p = .291$, 95% CI: -.224 to .742, $sr^2 = .005$). In other words, ERRT scores did not explain a significant amount of variance in ERIT-1 scores over and above SAT-CR scores, $R^2_{change} = .005$, $F_{change}(1,119) = 1.125$, $p = .291$. SAT-CR scores uniquely explained 36% of the variance in ERIT-1 scores, whereas ERRT scores only accounted for .5% of unique variance in ERIT-1 scores, after controlling for SAT-CR scores.

Clearly, ERRT scores do not have much utility in predicting ERIT-1 scores. Moreover, ERRT and SAT-CR scores have differential relationships with ERIT-1 scores, perhaps due to differences in reliability. For instance, reliability estimates for SAT-CR scores reported on The College Board's website are quite high (i.e., .90-.93), whereas interrater reliability estimates for the ERRT ranged from .78 to .82. In addition, unlike the ERRT, the ERIT-1 and SAT-CR use a selected-response item format. Differing measurement methods could be contributing to the differential relationships between ERIT-1, SAT-CR, and ERRT scores (Campbell and Fiske, 1959). That is, the ERIT-1 and SAT-CR might be more highly correlated because both use a selected-response, multiple-choice item format, compared to the constructed-response format of ERRT items.

RQ 4- Known Groups Validity Evidence

Given the ERIT-1 measures a lower level step in the ER process, students who possess a minimal amount of ER process skills should perform better on ERIT-1 items compared to a group of students with negligible ER skills. More specifically, freshmen that responded to ERIT-1 items in fall 2012 received no ER intervention and thus were expected to possess negligible ER skills, but freshmen that responded to the ERIT-1 in fall 2013 should possess minimal ER skills because they received a small dose of ER intervention prior to completing the test.

Twenty-six items were used to compare freshmen students from 2012 that had no ER intervention at JMU to freshmen students from 2013 that had a minimal dose of ER intervention during Orientation programming (i.e., 75-minute "It's Complicated: Ethical Reasoning in Action" activity) (See Figure 1). The 26 items were common across the two pilot versions of the ERIT-1 (i.e. the TERA and the TERB) and the current version of the

ERIT-1. As shown in Table 7, freshmen students responding to the TERA, TERB, and ERIT-1 versions of the test had similar demographic information including age, SAT scores, and ethnic backgrounds.

Recall, difficulty values represent the percentage of students that responded correctly to an item. In essence, difficulty values can be thought of as an indication of item "easiness;" when an item has a higher difficulty value, a greater percentage of students responded correctly to that item. Of the 26 common items across all three versions of the ERIT-1, 17 items had higher p values or "difficulty" values for the freshmen that received a minimal or "low" dose of ER intervention compared to freshmen that received no intervention (See Table 14). That is, a greater percentage of freshmen that received a low dose of ER intervention answered 17 of the 26 common items correctly compared to freshmen that received no intervention. For example, item 5 on the ERIT-1: "An employee decides he's not going to follow the dress code at work because he considers what he wears to be a matter of personal choice only" had a difficulty value of .52 for freshmen assessed in fall 2013 (i.e. freshmen that received a low dose of ER intervention). In contrast, item 5 on the ERIT-1 had a difficulty value of .32 for freshmen assessed in fall 2012 (i.e. freshmen that received no ER intervention). Thus, 52% of freshmen that received a low dose of ER intervention answered item 5 correctly, whereas only 32% of freshmen that received no ER intervention answered item 5 correctly. Given a larger percentage of freshmen expected to have some amount of ER abilities (i.e. freshmen that received a low dose of intervention) correctly responded to 17 out of the 26 common items compared to freshmen from 2012 that received no ER intervention, there is some support for known groups validity.

There were nine items for which a higher percentage of "no intervention" freshmen provided the correct answer compared to the freshmen that experienced the intervention (See Table 14). Items 10 and 11 had the largest discrepancy between difficulty values for the two groups of students. Compared to the freshmen that experienced the ER intervention, an additional 11% of students from the "no intervention" group answered items 10 and 11 correctly. Both items shared a common application area, but not a common key question. Items 10 and 11 were weakly correlated with the other items on the test (r = .230 and r = .162, respectively). Interestingly, over 20% of students in the "intervention" group indicated that Empathy was the correct response to items 10 and 11, perhaps due to the fact that both items referred to "elderly" neighbors or community members. Students could be relying too heavily on the reference to "elderly" individuals and thus incorrectly selecting Empathy as the most appropriate KQ. Given the references to "elderly" individuals in items 10 and 11 are tangential to the correct response for each item, they could easily be removed and the content of the items would remain the same.

Chapter V: Discussion

James Madison University's reaccreditation initiative focused on providing a model of ER and enhancing students' ability to engage in the ER process. JMU stakeholders created the ERIT-1 to assess students' lower level ER abilities (i.e., SLOs 2 & 3). As described in Chapter II, when evaluated using Benson's program of construct validation the ERIT-1 had several strengths including strong evidence aligned with Benson's (1998) substantive stage. For instance, recall that the Madison Collaborative worked with an ER content expert to articulate the conceptual and empirical definition of ER as a process. Then the Madison Collaborative operationally defined ER by transforming the theoretical conceptualization of ER into observable SLOs that could be assessed. An ER content expert facilitated item writing sessions to create items for the ERIT-1; items were written to measure a lower level step in the ER process (i.e., SLOs 2 & 3) (See Figure 2).

Although the ERIT-1 exhibited several strengths in terms of Benson's (1998) substantive stage, prior to this thesis, the ERIT-1 lacked evidence that aligned with the structural or external stages. Thus, to provide support for Benson's structural stage this thesis examined the factor structure of the ERIT-1and found evidence for an *essentially* unidimensional factor structure. Given a unidimensional factor structure, this thesis sought to provide evidence for Benson's external stage by examining the relationship between ERIT-1 scores, SAT-Critical Readings (SAT-CR) scores, and Ethical Reasoning Recall Test (ERRT) scores. The hypotheses regarding how the ERIT-1 should relate to other instruments were not fully supported. Specifically, the ERIT-1 correlated more highly with SAT-CR scores than expected and not as highly with another ER instrument,

the ERRT. Lastly, this thesis examined known groups validity evidence for ERIT-1 scores by comparing the performance of a group known to have minimal ER skills to a group that had negligible ER skills. Students who experienced a brief ER intervention at JMU performed better on 17 out of 26 common items compared to students that received no intervention.

Dimensionality and Scoring the ERIT-1

Confirmatory Factor Analysis (CFA) was used to inform ERIT-1 scoring. Recall that dimensionality is a fundamental aspect of test scores. The goal is to reduce down data to the lowest number of meaningful scores. In the ERIT-1's case, the data were somewhat consistent with the theory underlying the one-factor model, meaning that responses to the 42 items could be meaningfully represented with one score.

Nevertheless, the plausibility of a one-factor or "essentially unidimensional" model was challenged by the superior fit of the post hoc, bifactor "Liberty*" model. The items written to represent the Liberty KQ shared meaningful, non-negligible variance above and beyond the variance shared with the other non-Liberty items. But the issue for JMU stakeholders is that items on the ERIT-1 were written to represent one construct: lower level ability to engage in the ER process. The Liberty items, like all of the other items on the test, were created to get at an overarching "ER Process" factor. In short, all of the items should measure two lower level steps in the ER process, regardless of KQ or application area.

Thus, the CFA results have implications for test revisions. First, for the one-factor and bifactor models, items 7, 8, and 42 had the lowest factor loadings compared to the other items on the test (See Table 12). Moreover, items 7 and 42 did not have significant

factor loadings, which was not surprising because the highest observed correlation that any given item shared with item 42 was .193; the observed correlations between item 7 and the other items were also low. It may be appropriate to remove items 7, 8, and 42 from the ERIT-1; however, if item 7 is removed it should be replaced with an item that aligns with the Outcomes KQ and Civic application area to cover the breadth of the "ER process" construct (See Figure 3). Items 11, 39, and 41 had significant factor loadings for the one- factor model; however, their factor loadings were less than 0.30, indicating that less than 10% of the item's variance was explained by the "ER Process" factor. Stakeholders may consider revising or removing items 11, 39 and 41 based on the one-factor model.

Second, the "Liberty*" subfactor that emerged cannot be ignored; there is something salient that the Liberty items share over and above the overarching "ER Process" factor. If these items are multidimensional but they are modeled as unidimensional, the parameter estimates will be biased and any estimated relationships between the "ER process" factor, as measured by the ERIT-1, and other constructs will be biased (Reise, Bonifay, & Haviland, 2013). That is, if the items are multidimensional, reporting and using a single, total score for the ERIT-1 would be inappropriate because the items would share substantively meaningful common variance beyond the variance shared with the overarching "ER process" factor. Reporting a total score for items that are truly multidimensional would ignore meaningful, shared variance among the items that is due to something other than the "ER process" factor.

Limitations. There are several limitations associated with the bifactor model results for this study. One is that the bifactor model was specified based on LISREL

modification indices; therefore, it may be capitalizing on idiosyncrasies in the data and the results may not generalize to other samples (MacCallum et al., 1992). Interestingly, on average, the Liberty questions were the second most difficult items compared to the other KQs (See Table 8). It is possible that the difficulty of the Liberty items influenced their relationships with one another and the other items on the test. For instance, the Liberty items could share common variance simply because they were more difficult than other items, not because of a meaningful "Liberty*" subfactor. In other words, a substantively meaningful "Liberty*" subfactor may not exist.

Recall, the 829 students that responded to the ERIT-1 only experienced a 75-minute ER intervention during Orientation programming. Students are not expected to have gained an adequate understanding of the eight KQs through one 75-minute intervention. However, once students experience a stronger dose of ER intervention they should be able to describe the nuances of the Liberty KQ and what distinguishes Liberty items from Rights items. As students experience more ER interventions, the Liberty items should become less difficult. If the shared common variance among Liberty items was due to the fact that these items were some of the more difficult items on the test, then the shared common variance among the Liberty items that was observed in this study should be diminished in subsequent samples of students that experienced more ER intervention. Therefore, the bifactor model results might not replicate in a sample of students who receive higher "doses" of ER intervention.

Although there are some limitations associated with the bifactor model, there are implications for ERIT-1 scoring due to the "Liberty*" factor that emerged in this study. If the factor pattern coefficients associated with the six Liberty items and the "ER Process"

factor would have been larger than the factor pattern coefficients associated with the six Liberty items and the "Liberty*" factor, a total score for the test could easily be argued (summing together the remaining 36 items). However, this did not occur (See Table 12). As shown in Table 12, the six Liberty items were associated with the "ER Process" factor and the "Liberty*" factor fairly equally. Items 5 and 20 had the largest discrepancies between their relationship with the "ER Process" factor and the "Liberty*" factor; the discrepancy between the "ER Process" factor loading and "Liberty*" factor loading for the remaining four Liberty items was less than |.07|. It can be argued that reporting a total score for the ERIT-1 is inappropriate given the emergence of the "Liberty*" subfactor. Yet, applying the bifactor model would make the ERIT-1 more difficult to score. Given scoring difficulty, one may advocate for removing the Liberty items. However, if the Liberty items were excluded from the test, the ERIT-1 would not cover the breadth of the ER domain (See Figure 3). The emergence of the Liberty* subfactor suggests directions for future research concerning the ERIT-1.

Future research. To inform stakeholders about scoring the ERIT-1, future research should attempt to replicate the findings of this study. If the findings replicate, researchers should compare the one-factor and bifactor models when predicting several different, appropriate outcomes related to ERIT-1 scores (Yost & Finney, in press). If the "ER Process" factor for the one-factor model and the "ER Process" factor for the bifactor model differentially predict the various outcomes, then the ERIT-1 should not be scored as a unidimensional instrument (Reise et al., 2013). Alternatively, if the "ER Process" factor for the one-factor and bifactor models have similar relationships with the various outcomes, this finding would provide further support for the more parsimonious, one-

factor model. Using the bifactor model, the relationship between the "Liberty*" subfactor and the various outcomes should also be estimated. If the "ER Process" factor and the "Liberty*" subfactor related to the various outcomes in a similar way, this finding would provide additional support for a unidimensional structure; however, if they related differentially the bifactor model should be further investigated. One relevant outcome variable that could be used in future research is "openness to experience" (Costa & McCrae, 1992). Students who have more of what Costa and McCrae (1992) defined as "openness to experience" are expected to be more prone to exploring the different perspectives that constitute the KQs and this "openness" should correlate similarly with the "ER Process" factor than the "Liberty*" subfactor given the ERIT-1 is unidimensional. Moreover, the overarching "ER Process" factor for the one-factor and bifactor models should have similar relationships with "openness."

Another appropriate outcome variable could be moral reasoning. Recall, ER and moral reasoning are closely linked in theory. Moral reasoning typically refers to an empirical behavior, or how a person *actually* reasons, and ER refers to how a person *should* reason. The "ER Process" factor for the one-factor and bifactor models should have similar relationships with moral reasoning. That is, the "ER Process" factor for the one-factor and bifactor models should *not* differentially predict moral reasoning. Using the bifactor model, the correlation between moral reasoning and the "ER Process" factor should be similar to the correlation between moral reasoning and the "Liberty*" subfactor; this finding would provide evidence to further support the one-factor model.

Future research should also estimate the one-factor and bifactor models using students that have experienced the full extent of the Madison Collaborative's ER

interventions. It is plausible that the internal structure of the ERIT-1 could be different for a group of students that received all of the ER interventions and thus have a deeper understanding of the KQ framework. For instance, as students learn more about the eight KQs, they should better understand the subtle nuances associated with the Liberty KQ. Thus, Liberty items should become less difficult for students. If the shared variance among Liberty items was influenced by item difficulty, then it could be diminished when students that experienced more ER intervention (i.e., have a more in-depth understanding of the KQs) are sampled. In that case, the factor structure of the ERIT-1 could become more unidimensional when sampling students that experienced more ER interventions.

In addition, future research should investigate the eight testlet items on the ERIT-1. That is, items 43 through 50 form two testlets comprised of four items each; items 43 through 46 are based on one scenario and items 47 through 50 are based on a different scenario. This thesis focused on the 42 non-testlet ERIT-1 items, yet items 43 through 50 still need to be examined. Before stakeholders can make decisions about scoring the 50-item ERIT-1, the dimensionality of the test should be reassessed including the eight testlet items. Until then, the "essentially unidimensional" structure can only be applied to the first 42 items on the ERIT-1.

Initial External Validity Evidence for ERIT-1 Scores

In order to build a body of validity evidence for ERIT-1 scores, one purpose of this thesis was to collect convergent, divergent, and known groups validity evidence. Prior to conducting this thesis, a major weakness of the ERIT-1 was that it lacked evidence to support Benson's (1998) external stage of construct validation. Gathering validity evidence for ERIT-1 scores will help stakeholders understand the meaningfulness

of ERIT-1 test scores, and hopefully contribute to more accurate inferences about students' ability to engage in the ER process.

To collect such evidence, the relationship between ERIT-1 and ERRT scores was estimated in an attempt to provide convergent validity evidence. In addition, the relationship between ERIT-1 scores and SAT-CR scores was estimated in an attempt to provide divergent validity evidence for ERIT-1 scores. To provide known groups validity evidence, ERIT-1 test performance of a group known to possess a small amount of ER process skills was compared to test performance of a group that possessed no ER process skills as measured by the ERIT-1.

Concerning convergent validity evidence, ERIT-1 scores were statistically significantly positively correlated with ERRT scores. Yet, the statistically significant correlation between ERIT-1 and ERRT scores was not practically significant; ERRT scores accounted for only 6.6% of the variability in ERIT-1 scores. ERIT-1 scores were statistically significantly more strongly correlated with SAT-CR scores than with ERRT scores. I hypothesized that ERIT-1 scores should have been more strongly correlated with ERRT scores because both tests purport to measure students' abilities to engage in lower level steps of the ER process. However, an ER content expert commented that he expected ERIT-1 scores to be more strongly correlated with SAT-CR scores than ERRT scores. Although, the hierarchical nature of the SLOs suggest that students should achieve the skills assessed by the ERRT (i.e. SLO 1) before they can achieve the skills assessed by the ERIT-1 (i.e. SLOs 2 & 3), these two instruments do not measure the same *kind* of ER skills. For example, a student that can successfully recall from memory the eight KQs and define each one (i.e., SLO 1) may not be able to successfully identify the

relationship between a KQ and a given rationale (i.e., SLOs 2 & 3). Due to the fact that the ERIT-1 and ERRT measure two different kinds of skill, the initial hypothesis about the strength of their relationship was overstated. That is, I should not have anticipated such a strong relationship between ERIT-1 and ERRT scores. Rather, I should have expected ERIT-1 and ERRT scores to share a positive relationship, and given that they did, these findings provided initial convergent validity evidence.

Concerning divergent validity evidence for ERIT-1 scores, over 80% of the variability in ERIT-1 scores could not be explained by SAT-CR scores. ERIT-1 scores likely measure a construct that is substantively distinct from verbal proficiency. This finding provides some initial divergent validity evidence for ERIT-1 scores. Also, this finding addresses the concerns of one faculty member that criticized the ERIT-1 for being "merely a vocabulary test" (David McGraw, personal communication, October 23, 2012).

Lastly, this thesis provided some evidence of known groups validity. Theory suggests that the freshmen who experienced a small dose of ER intervention should perform better on the ERIT-1 compared to freshmen that did not experience any ER interventions at JMU. As shown in Table 14, a larger percentage of freshmen expected to have some amount of ER abilities (i.e. freshmen assessed in 2013 that received a low dose of intervention) correctly responded to 17 items compared to freshmen assessed in 2012 that received no ER intervention. Comparing the difficulty values for the 26 common items across the pilot versions of the ERIT-1 (i.e. the TERA and the TERB) and the current version of the ERIT-1 demonstrated that 17 of the 26 common items were easier for freshmen that experienced the ER intervention than for freshmen that experienced no intervention. These findings aligned with the Madison Collaborative's

theory suggesting that the majority of common items on the ERIT-1 were easier for the group of students that was expected to possess a small amount of ER abilities.

Limitations. The estimated correlations between ERIT-1 and ERRT scores could be attenuated by measurement error or unreliability in test scores. However, reliability estimates for the ERIT-1 and the ERRT were similar; scores from both instruments demonstrated adequate reliability (α = .787 and average interrater reliability = .798, respectively). Thus, the weak correlation between ERIT-1 and ERRT scores is probably not due to measurement error associated with the ERRT. The correlation between ERIT-1 and ERRT scores disattenuated for unreliability was moderate, $r(140) = .324$ (Cohen, .1992).

In addition, a limitation of the known groups validity evidence is that the group expected to possess a "small" amount of ER skills only experienced a 75-minute ER intervention during Orientation programming. It is unlikely that these students were able to gain very many ER process skills during this short intervention; the 75-minute "dose" of ER intervention might have been too small to evidence tangible ER process abilities. There are more comprehensive interventions that future cohorts of students will experience such as an online, year-long ER course that all freshmen students must complete during their first year at JMU (The Madison Collaborative: Ethical Reasoning in Action, 2013, p. 36).

Future research. As more external validity studies are conducted on these two instruments, the theorized hierarchy of the SLOs should be evaluated and further convergent validity evidence should be collected. Perhaps the hierarchical "step" between SLO 1 and SLOs 2 and 3 is not as distinct as the Madison Collaborative theorized. That

is, the correlation between ERIT-1 and ERRT scores suggested that SLOs 1, 2, and 3 might all represent one lower level step in the ER process. Students might not necessarily have to master SLO 1 before they can achieve SLO 2 or 3. Furthermore, students that perform well on the ERRT might not necessarily perform well on the ERIT-1 because these instruments measure different kinds of ER skills. Students are given a brief definition of the 8 KQs on the ERIT-1; thus, they do not have to recall the KQs from memory in order to do well on the ERIT-1. Given the ERIT-1 and ERRT measure different kinds of ER skills, the hypothesized relationship between ERIT-1 and ERRT scores was overstated. To collect convergent validity evidence, future research should estimate the correlation between ERIT-1 scores and scores from other ER instruments; these instruments should measure ER skills that are more similar to the skills measured by the ERIT-1.

Concerning the divergent validity evidence, the presence of a method effect could explain the moderate correlation between ERIT-1 and SAT-CR scores. As Campbell and Fiske (1959) described, the extent to which method variance influences scores, the scores are invalid. Ideally, there is minimal method variance (i.e., there is no method effect). Given the ERIT-1 and SAT use the same measurement method (i.e., selected-response), the moderate observed correlations between ERIT-1 and SAT-CR scores might be a function of method variance that is common among the ERIT-1 and the SAT but not common to the ERRT (Campbell & Fiske, 1959).

To apply Campbell and Fiske's multitrait-multimethod matrix to the ERIT-1, ERRT, and SAT-CR, new versions of the ERIT-1 and SAT-CR that use a constructed-response item format would need to be created. Also, a new form of the ERRT that uses a

selected-response item format must be created. The matrix would consist of two separate measurement methods (i.e., selected-response and constructed-response format), in addition to three different traits: ER- KQ recall, ER-KQ application, and Verbal Proficiency. If there were a method effect of item format, in terms of Campbell and Fiske's (1959) multitrait-multimethod matrix, we would expect to see some of the highest correlations among different traits (i.e., ER- KQ recall, ER-KQ application, and Verbal Proficiency) from the same method. More concretely, the highest observed correlations in the matrix would be among ER-KQ recall, ER-KQ application, and Verbal Proficiency (i.e., different traits) from the same measurement method (i.e., selected-response or constructed-response). Theoretically, new versions of the ERIT-1 and SAT-CR could be created to have constructed-response item formats. However, it does not make sense to create a new version of the ERRT that is selected-response. Recall the ERRT measures SLO 1 (i.e., students' ability to state, from memory, all eight KQs), which requires a constructed response item format. If ERRT items were converted to a selected-response format, the ERRT would no longer align with SLO 1. Thus, it would be difficult to empirically construct a multitrait-multimethod matrix for the ERIT-1, ERRT, and SAT-CR.

Given it would be impractical to construct a multitrait-multimethod matrix, future research should focus on other means of collecting convergent validity evidence for ERIT-1 scores. As discussed previously, the hypothesized relationship between ERIT-1 and ERRT scores was overstated because the ERIT-1 and ERRT measure different kinds of ER abilities. Therefore, future research should focus on the relationship between

ERIT-1 scores and scores from other ER instruments that measure the same kind of ER skills as the ERIT-1.

In addition, researchers need to collect additional known groups validity evidence using different groups. For example, the Madison Collaborative should compare groups that received higher "doses" of ER intervention. Colleagues from the JMU community that will be training faculty members to integrate the 8 KQs into their classroom curricula are one example of a group known to possess ER skills. Presumably, these members of the JMU community had a "strong dose" of instruction about the ER process because they will be teaching the 8 KQ framework to other faculty members. Given their in-depth knowledge of the KQs and the Madison Collaborative, this group would be expected to perform better on the ERIT-1 than students that received a "medium," "low," or "no dose" of ER intervention.

Another example of a group the Madison Collaborative should investigate to collect known groups validity evidence is a group of students that took an ER course taught by Professor William Hawk during the spring 2013 or a future cohort of students that take a similar course. They had a "medium dose" of instruction about the KQ framework and the ER process as defined by the Madison Collaborative. Given their previous instruction, this group would be expected to perform better on the ERIT-1 compared to a group of students that did not take Dr. Hawk's course (i.e., a group of students that received a "low" or "no dose" of ER intervention). Investigating groups with varying degrees of ER skills, similar to the groups just described, will further bolster the known groups validity evidence for ERIT-1 scores.

Implications and Conclusions.

Although the ERIT-1 is still in its infancy in terms of empirical evaluation, so far, it has demonstrated great potential for assessing Madison Collaborative ER student learning outcomes. That is, the ERIT-1 has a solid theoretical foundation developed by ER content expert Dr. William Hawk and other members of the Madison Collaborative. Moreover, the concept of ER, as measured by the ERIT-1, has sufficient evidence aligned with Benson's (1998) substantive stage of construct validation. This thesis contributed initial evidence aligned with Benson's (1998) structural and external stages. The implication of this work is that the ERIT-1 can be considered an adequate measure to assess students' ER skills at JMU, and total scores can be reported and analyzed.

Administrators and stakeholders at JMU focused their reaccreditation efforts on the need "to elevate ethical reasoning as a priority for undergraduate student learning" (The Madison Collaborative: Ethical Reasoning in Action, 2013, p. 1). Recall, the overarching goal of the Madison Collaborative initiative was to prepare *all* students to be "informed and enlightened citizens" that actively engage in the ER process. To achieve this goal, the Madison Collaborative provided a KQ framework for teaching and assessing ER abilities; designed campus-wide interventions to directly influence every student; and created assessment tools to gauge student learning and evaluate the effectiveness of ER interventions.

Given the widespread prevalence of ethical dilemmas, our society needs citizens that can engage in the process of ER. While many institutions have embraced the concept of ethical reasoning, few have implemented wide-reaching programs to develop it. As Kohlberg (1977) asserted, cultivating ER skills requires effortful development.

Furthermore, enhancing ER skills across a variety of disciplines requires interventions that directly impact every student and psychometrically sound instruments that capture gains in students' ER abilities. JMU has made commendable strides toward ensuring that all students are educated about the ER process. Through the creation and development of the ERIT-1, stakeholders at JMU are one step closer to capturing the impact of this noble pursuit.

Appendix A

Item Analyses Information for the TERA and TERB

Keep	Version	Item	Difficulty	Standard Deviation	Discrimination	Alpha if deleted	Comments
X	A	1	0.52	0.50	0.29	0.757	
	A	2	0.80	0.40	0.21	0.760	Deleting would decrease overall alpha. I think it could stay but would need some revision (.20<=D<=.29). ~80% of people are responding correctly, so could it be too easy? Between this Occupation item (i2A) and the following one (i2B) they both had the same discrimination so it was hard to pick between them. We already have one Occupation item (i37B) which has a higher discrimination than both of these two so I can only pick one of these two. I went with i2B because less people got it correct on average but it is still discriminating just as well as i2A.
	A	3	0.95	0.22	0.13	0.762	Not doing the best job at discriminating. ~95% are answering correctly. It is likely too easy. Should be deleted or completely revised (D<=.19). We already have two Professional Items and they both had better discrimination so I just went with the two Professional Items that had the higher discrimination and I decided not to keep this one
X	A	4	0.35	0.48	0.21	0.760	
X	A	5	0.88	0.33	0.17	0.761	
	A	6	0.19	0.39	0.06	0.765	Only 19% of people are getting it right. Could it be confusing? Or ambiguous? It is probably too difficult. Should be eliminated or completely revised (D<=.19). Most frequent answer was Liberty. Second most frequent answer was Responsibilities. Since the item involves running for office maybe people associate that with Liberty or Responsibilities as in civic responsibilities? We already have two Civic Items and they both had better discrimination so I just retained the two Civic items that had the higher discrimination.

	A	7	0.19	0.39	0.18	0.761	Overall alpha wouldn't change much if this item were deleted. Only ~19% of people are getting it right. Could it be confusing? Or ambiguous? It is likely too difficult. Should be eliminated or completely revised (D<=.19). Most frequent answer was Liberty. Second most frequent answer was Responsibilities. I could see why students are getting confused here. Maybe they thought that saying no to burning the flag related to "principles of liberty"
X	A	8	0.67	0.47	0.29	0.757	
X	A	9	0.24	0.42	0.31	0.756	
	A	10	0.74	0.44	0.23	0.759	Deleting would decrease overall alpha. It needs some revision (.20<=D<=.29). ~74% of people are responding correctly, so maybe it should be more difficult. We already have two Civic items and they both had slightly better discrimination so I decided not to retain this item
X	A	11	0.76	0.43	0.18	0.761	
X	A	12	0.66	0.47	0.24	0.759	
X	A	13	0.31	0.46	0.25	0.759	
	A	14	0.88	0.33	0.05	0.765	Not doing the best job at discriminating. ~88% are answering correctly. Item might be too easy. Should be deleted or completely revised (D<=.19). Students who got this question wrong by answering Liberty, still did almost just as well on their average score on the test. We already have two Personal items and they both had better discrimination so I decided not to keep this one.
X	A	15	0.48	0.50	0.31	0.756	
	A	16	0.64	0.48	0.22	0.760	Item needs some revision (.20<=D<=.29). Item is not functioning terribly but we have other Self items that are functioning better, so to balance content we are not retaining this item
X	A	17	0.66	0.47	0.27	0.757	
	A	18	0.47	0.50	0.21	0.760	Item needs some revision (.20<=D<=.29). We already have two Occupational items and they both had slightly better discrimination so I decided not to keep this one

	A	19	0.64	0.48	0.22	0.760	Item needs some revision (.20<=D<=.29). I didn't keep this one because item 47 is also Personal domain and it's part of a testlest so we have to keep it and we already have item 15 which was also Personal. Compared to item 15, this item was easier and did not discriminate as well, so we retained 15 and dropped this item
	A	20	0.91	0.28	0.17	0.761	Item needs some revision (.20<=D<=.29). 90% of students are responding correctly. Item is just too easy.
X	A	21	0.87	0.34	0.23	0.759	
X	A	22	0.91	0.28	0.23	0.760	
	A	23	0.33	0.47	0.11	0.764	Only 1/3 of people are getting it right. Could it be confusing? Or ambiguous? It is probably too difficult. Should be eliminated or completely revised (D<=.19). Second most frequent answer was Self, third most frequent was responsibilities followed closely by Rights. People who answered "Outcomes" got higher scores on test on average. Maybe it's confusing? It seems like the item could be referring to Self or Rights instead of the correct answer, Liberty. We already have two Personal items and they both had better discrimination so I decided not to keep this one.
X	A	24	0.62	0.49	0.35	0.754	
	A	25	0.95	0.21	0.12	0.763	Item is not discriminating well. 95% are answering correctly. Item is way too easy. It should be deleted or completely revised (D<=.19). Students who answered Authority to this item (got this item wrong) had very slightly higher scores on the test than people who got this question correct, which is NOT what we want at all, so this item was dropped.
	A	26	0.22	0.41	0.04	0.766	Item is not discriminating well. Should be deleted or completely revised (D<=.19). Students who answered Liberty (the wrong answer) to this item on average scored higher on the test, which is NOT what we want at all, so this item was dropped.
X	A	27	0.76	0.43	0.22	0.760	
	A	28	0.87	0.33	0.21	0.760	Item needs some revision (.20<=D<=.29). ~87% of people are responding correctly, so item might be too easy. I didn't keep this item because items 46 and 48 are also Personal domain and they are part of a testlest so we have to keep them. We had two Personal items which was enough so I decided not to keep this item

X	A	29	0.98	0.16	0.23	0.761	
X	A	30	0.53	0.50	0.32	0.755	
	A	31	0.50	0.50	0.16	0.762	Item is not doing the best job at discriminating. It should be deleted or completely revised (D<=.19). Second most frequent answer was Fairness. I guess if Karen was "balancing the interests" of the people who might be offended I could see how someone might think of this as "fairness" instead of outcomes. I didn't keep this one because item 45 is also Professional domain and it's part of a testlest so we have to keep it and we already kept item 24 which was also Professional and was a better item. Thus, we didn't need more professional/outcomes items
	A	32	0.38	0.49	0.18	0.762	Only about 38% of students are getting it right. Might be too difficult. Item should be eliminated or completely revised (.20<=D<=.29). We already have two Occupational items and they both had better discrimination so I decided to retain the two Occupational items that had the higher discrimination values
X	A	33	0.35	0.48	0.23	0.759	
X	A	34	0.71	0.45	0.30	0.757	
X	A	35	0.78	0.42	0.27	0.758	
	A	36	0.55	0.50	0.10	0.765	Not doing the best job at discriminating. Item should be deleted or completely revised (D<=.19). Second most frequent answer is outcomes. I could see why people might have said outcomes. Because it talked about developing skills. I think this item was confusing. It confused me at least.
X	A	37	0.87	0.34	0.32	0.757	
	A	38	0.59	0.49	0.23	0.759	Item needs some revision (.20<=D<=.29). I didn't keep this one because item 47 is also Civic domain and it's part of a testlest so we have to keep it and we already kept item 38 which was also Civic and was performing well. Thus, we did not need another Civic item.
X	A	39	0.09	0.28	0.11	0.763	

X	A	40	0.57	0.50	0.21	0.760	
	A	41	0.69	0.46	0.32	0.755	Item is discriminating fairly well. Little or no revision would be required (.3<=D<=.39). This item was deleted because it was part of first testlet on version A and we decided that the second testlet was better for A and we can't break up the testlets so although this one was doing fine it was lumped with the testlet that didn't do as well as the other one on form A so we can't use this one.
	A	42	0.74	0.44	0.26	0.758	Item needs some revision (.20<=D<=.29). ~74% of people are responding correctly, so could it be too easy? This item was part of first testlet on version A and we decided that the second testlet was better for A and we can't break up the testlets so although this one was doing fine it was lumped with the testlet that didn't do as well as the other one on form A so we can't use this one.
	A	43	0.66	0.47	0.08	0.765	Not doing the best job at discriminating. Item should be deleted or completely revised (D<=.19). Second most frequent answer was Empathy. I think it is pretty clear that it is outcomes. 66% got it correct so most people got what it was trying to convey. However, I can see why someone who might not really understand the 8 KQs would put empathy because it talks about "how you would respond if you cared deeply about people involved" and if his family cared deeply about him then they would suffer if he died. This was from the first testlet set in version A and we decided to keep the second testlet and we can't break up the testlet so this item ended up in the testlet we are not keeping so we can't keep this item.
	A	44	0.60	0.49	0.24	0.759	Item needs some revision (.20<=D<=.29). This was from the first testlet set in version A and we decided to keep the second testlet and we can't break up the testlet so this item ended up in the testlet we are not keeping so we can't keep this item.
X	A	45	0.80	0.40	0.29	0.757	
X	A	46	0.64	0.48	0.27	0.758	
X	A	47	0.73	0.44	0.38	0.753	
X	A	48	0.40	0.49	0.28	0.757	

	B	1	0.25	0.43	0.04	0.694	Overall alpha wouldn't change much if this item were deleted. Only 25% of students are getting it right. Could it be confusing? It's probably too difficult. Item should be eliminated or completely revised (D<=.19). The most frequent answer was Responsibilities, NOT the correct answer which was authority. I could see where people might think that Jessica was "obligated" to tell on her coworker. The majority of people got this wrong. People who answered Liberty to this item on average scored higher on the test. than people who provided the correct answer, which was authority
X	B	2	0.77	0.42	0.20	0.685	
	B	3	0.97	0.16	0.01	0.692	Almost everyone is getting the item correct (97%). It's doing a poor job at discriminating because basically everyone is getting it right. Deleting it won't have much of an effect on alpha. Item should be eliminated (D<=.19). Students who answered Self to this item on average scored higher on the test than people who answered this item correctly. It's doing a terrible job discriminating. The correct answer might be too obvious.
X	B	4	0.38	0.49	0.32	0.677	
X	B	5	0.48	0.50	0.21	0.684	
	B	6	0.30	0.46	0.01	0.696	Only about 1/3 of people are getting it right. Might be too hard. Item should be eliminated or completely revised (D<=.19). Many people also said the answer was "Authority" or "Responsibilities". School boards and parents could be seen as authority figures. Maybe the scenario is unclear because the correct answer is Outcomes and not Authority or Responsibilities.
	B	7	0.52	0.50	0.20	0.685	Not doing the best job at discriminating. Item needs some revision (.20<=D<=.29). The second most frequent answer was "Liberty". I am not sure how people got "Liberty" out of this one. I don't see anything about liberty or "personal autonomy" in this item. About 52% of people got it right. Maybe the quotation "I am a citizen of democracy" confused people and made them associate that with "Liberty" instead of "Responsibilities"?
X	B	8	0.80	0.40	0.25	0.683	
X	B	9	0.27	0.44	0.13	0.689	

X	B	10	0.71	0.45	0.24	0.683	
X	B	11	0.85	0.35	0.17	0.687	
	B	12	0.72	0.45	0.05	0.694	Not doing the best job at discriminating. Seems like a lot of people are answering it correctly (72%). Might be too easy. Item should be completely revised or eliminated (D<=.19). The majority of people got it right. Maybe because the scenario is about voting that is priming or cluing people in to "rights". We already have more than enough Civic items and they all had better discrimination so I decided to get rid of this one.
X	B	13	0.47	0.50	0.27	0.681	
	B	14	0.31	0.46	0.02	0.696	Only about 1/3 of students are getting it right. Might be too hard. Item should be eliminated or completely revised (D<=.19). Many people said the answer was "Fairness". Maybe they got confused because allowing health care benefits for same sex partners could be considered fair. Item might be confusing? Also, a similar pattern happened with item 39. The correct answer was rights but it had to do with same sex partnership/marriage and for both items many students answered "Fairness" instead of Rights.
	B	15	0.61	0.49	0.19	0.686	Item needs some revision (D<=.20). I didn't keep this item because item 47 is also Personal domain and it's part of a testlest so we have to keep it and we already kept item 15 which was also Personal so we had two enough Personal items that are functioning better.
	B	16	0.81	0.39	0.02	0.694	Not doing the best job at discriminating. 80% are answering correctly. Maybe item should it be more difficult. Item should be deleted or completely revised (D<=.19). The majority of people are responding correctly. The correct answer might be too obvious. We already have more than enough Personal items and they all had better discrimination so I dropped this one.
	B	17	0.82	0.39	0.19	0.686	Item needs some revision (D<=.20). 82% of students are responding correctly, so could it be too easy.
X	B	18	0.78	0.41	0.23	0.684	
X	B	19	0.48	0.50	0.26	0.682	

	B	20	0.97	0.18	0.10	0.690	Almost everyone is getting the item correct. It's doing a poor job at discriminating because basically everyone is getting it right. Deleting it won't have much of an effect on alpha. Item should be eliminated (D<=.19). People are only endorsing three of the 8 response options. The correct answer is probably too obvious
	B	21	0.78	0.42	0.16	0.687	Item needs revision (D<=.19). 78% of students are responding correctly, so could it be too easy. The second most frequent answer was "Self", but the majority if people got it right. I think it pretty clearly points to "Empathy" rather than an ideal "Self". But then again if he does not like to see animals suffer, helping animals could be an action that would help him become his ideal self? If you thought about it like that, I could see this one being a tricky choice between "Empathy" and "Self"
X	B	22	0.78	0.41	0.18	0.686	
	B	23	0.26	0.44	0.14	0.689	Overall alpha wouldn't change much if this item were deleted. Only 26% of students are getting it right. Item might be confusing or ambiguous? Also, item might be too hard. Item should be eliminated or completely revised (D<=.19). Students who answered Empathy to this item on average scored higher on the test than people who answered this item correctly. The second most frequent answer was "Outcomes". I can see why they might think "Outcomes" because the item talks about Jennifer being "reprimanded later on because of an action she did now- speaking up in meetings". Maybe the reprimanded part is confusing students.
X	B	24	0.95	0.22	0.23	0.687	
X	B	25	0.60	0.49	0.23	0.683	
	B	26	0.30	0.46	0.09	0.692	Only about 1/3 of people are getting it right. Might be too hard. Item should be eliminated or completely revised (D<=.19). The most frequent answer was "Empathy", NOT the correct answer, "Self". Not many people got this one correct. Maybe it should be more explicit that the author is creating the bill for reasons related to his ideal self/personality. Maybe because the bill is about the forest people confuse that with "empathy". Could the bill be about something that people would not link to empathy like the preservation of forests/nature? We actually need another Civic item here! However, this item is doing so poorly, discrimination is <.15, so it's not really good enough for us to keep this item even though we need another Civic item.

X	B	27	0.44	0.50	0.28	0.680	
	B	28	0.34	0.48	0.18	0.686	~34% of students are getting it correct. Might be too hard or confusing? Item should be eliminated (D<=.19). A slightly higher frequency responded that the correct answer was "Self". The majority of people got it wrong and there must have been some confusion about the differences between "Self" and “empathy". I would think that an example involving bullying would clue people in to "Empathy". Maybe because "Cora walked away" people thought that was selfish and she did not want to get bullied again so she was protecting herself? But that line of reasoning does not fit with the definition for self that was provided.
X	B	29	0.94	0.24	0.22	0.686	
	B	30	0.63	0.48	-0.03	0.699	Negative relationship. Eliminate or totally revise. Many students said the answer was "Liberty". Maybe people were confused by what a union really is? We actually need another Occupational item here! However, this item is doing so poorly, D is <.15, so it's not really good enough for us to keep this item even though we need another Occupational item.
	B	31	0.43	0.50	0.14	0.689	Not doing the best job at discriminating. Item should be deleted or completely revised (D<=.19). The second most frequent answer was "Authority". Maybe this item is unclear. I could see why a person might think of authority instead of Outcomes because the scenario mentions the university president. Maybe we should not have used an authority figure like the president because maybe that is causing confusion.
X	B	32	0.36	0.48	0.26	0.682	
X	B	33	0.47	0.50	0.31	0.678	
	B	34	0.53	0.50	0.21	0.685	Not doing the best job at discriminating. Item needs some revision (.20<=D<=.29). The second most frequent answer was "Outcomes". I think that "transportation experts" should clue students in to saying "Authority" is the right answer. It seems that some people thought of the traffic re-routing as an "outcomes" KQ? I guess people can see the short/long-term outcomes of re-routing traffic? Maybe the reference to authority was not explicit enough?

	B	35	0.64	0.48	0.19	0.686	Item needs some revision (D<=.20). We already have two Personal Items and they both had better discrimination so I decided not to keep this one.
X	B	36	0.66	0.47	0.19	0.686	
X	B	37	0.74	0.44	0.30	0.679	
X	B	38	0.60	0.49	0.26	0.681	
	B	39	0.28	0.45	-0.01	0.697	Negative relationship. Item should be dropped. Many people said the answer was "Fairness". Maybe they got confused because allowing same sex marriage could be considered fair. It might be confusing to examinees?
	B	40	0.40	0.49	0.04	0.695	Not doing the best job at discriminating. Item should be deleted or completely revised (D<=.19). The second most frequent answer was "Empathy". Maybe this item should was not clear enough. I could see why a person might think of empathy instead of fairness. Maybe it should be more apparent that this item is talking about fairness and NOT empathy. Although we really need another Civic item because we only have one right now, this item is just not functioning well enough to warrant keeping it.
	B	41	0.49	0.50	0.41	0.672	Item is discriminating well. Little or no revision would be required. However, this item was dropped because it was part of first testlet on version B and we decided that the second testlet was better from version B and we can't break up the testlets so although this one was doing fine it was lumped with the testlet that didn't do as well as the other one on version B so we can't use this item.
	B	42	0.94	0.24	0.18	0.687	Almost everyone is getting the item correct. It's doing a poor job at discriminating because basically everyone is getting it right. Item should be eliminated or completely revised (D<=.19). The correct answer is probably too obvious. This was from the first testlet set in version B and we decided to keep the second testlet and we can't break up the testlet so this item ended up in the testlet we are not keeping.

	B	43	0.74	0.44	0.14	0.689	Not doing the best job at discriminating. Item should be deleted or completely revised (D<=.19). People who answered Outcomes to this item on average scored higher on the test than people who answered this item correctly. The second most frequent answer was "Empathy". Maybe people responded this way to this item because of the phrase "stopping to see If anyone needs help". Maybe the "feeling obligated as a human being" part should be more emphasized or explicit and the "caring about anybody needing help" should be eliminated or shortened and made less apparent? This was from the first testlet set in version B and we decided to keep the second testlet and we can't break up the testlet so this item ended up in the testlet we are not keeping.
	B	44	0.84	0.37	0.10	0.691	Not doing the best job at discriminating. 84% are answering correctly. Item might be too easy. Item should be deleted or completely revised (D<=.19). Correct answer might have been too obvious? This was from the first testlet set in version B and we decided to keep the second testlet and we can't break up the testlet so this item ended up in the testlet we are not keeping.
X	B	45	0.91	0.29	0.30	0.683	
X	B	46	0.94	0.25	0.34	0.682	
X	B	47	0.90	0.31	0.22	0.685	
X	B	48	0.68	0.47	0.19	0.686	

References

Albaum, G., & Peterson, R. A. (2006). Ethical attitudes of future business leaders: Do they vary by gender and religiosity? *Business and Society, 45*(3), 300-321.

Alschuler, A. S., & Blimling, G. S. (1995). Curbing epidemic cheating through systemic change, *College Teaching, 43* (4), 123–126.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999*). Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Andreason, A. W. (1976). *The effects of social responsibility, moral judgment, and conformity on helping behavior*. (Doctoral dissertation). Provo, UT: Brigham Young University.

Aristotle. *Nicomachean ethics*. Translated by C. C. W. Taylor. Oxford, NY: Oxford University Press, Inc. 2006.

Association of American Colleges and Universities. (2013). It takes more than a major: Employer priorities for college learning and student success. *Liberal Education, 99*(2). Retrieved from http://www.aacu.org/liberaleducation/le-sp13/hartresearchassociates.cfm.

Augustine, N. R. (2013). One cannot live by equations alone: Education for life and work in the twenty-first century. *Liberal Education, 99*(2). Retrieved from http://www.aacu.org/liberaleducation/le-sp13/augustine.cfm.

Bailey, C. D. (2011). Does the defining issues test measure ethical judgment ability or political position? *The Journal of Social Psychology, 151*(3), 314-330.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*(1), 10-22.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psycholigcal Bulletin, 88*(1), 1-45.

Cahn, S. T., Markie, P. (2002). *Ethics: History, theory, and contemporary issues* (2nd ed.). Oxford, NY: Oxford University Press, Inc.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment: Theoretical foundations and research validations*. Cambridge, MA: Cambridge University Press, Inc.

Colby, A., Kohlberg, L., Gibbs, J., Lieberman, M., Fischer, K., & Saltzstein, H. D. (1983). A longitudinal study of moral judgment. *Monographs of the Society for Research in Child Development, 48*(1), 1-124.

Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: the NEO personality inventory. *Psychological Assessment, 4*(1), 5-13.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Dalton, J., & Crosby, P. (2011). Core values and commitments in college: The surprising return to ethics and character in undergraduate education. *Journal of College and Character, 12*(2), 1-4.

Davison, M. L., & Robbins, S. (1978). The reliability and validity of objective indices of moral development. *Applied Psychological Measurement, 2*(3), 391-403.

Denis, L. (2012). Kant and hume on morality. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Stanford, CA: The Metaphysics Research Lab. Retrieved from http://plato.stanford.edu/archives/fall2012/entries/kant-hume-morality/>.

Dolinger, S. J., & LaMartina, A. K. (1998). A note on moral reasoning and the five-factor model. *Journal of Social Behavior and Personality, 13*(1), 349-358.

Elm, D. R., & Weber, J. (1994). Measuring moral judgment: The moral judgment interview or the defining issues test?. *Journal of Business Ethics, 13*(5), 341-355.

Emler, N., Renwick, S., & Malone, B. (1983). The relationship between moral reasoning and political orientation. *Journal of Personality and Social Psychology, 45*, 1073-1080.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 109*(42), 17028-17033.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation models. In G.R. Hancock & R.O. Mueller (Eds.). *A second course in structural equation modeling*. Greenwich, CT: Information Age Publishing, Inc.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1*, 1-32.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5- 15.

Freeman, S. (2002). Rawls, John. In *Routledge Encyclopedia of Philosophy Online*. London: Routledge. Retrieved from http://www.rep.routledge.com/article/S091.

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.

Guyer, P. (2004). Kant, Immanuel. In *Routledge Encyclopedia of Philosophy Online*. London: Routledge. Retrieved from http://www.rep.routledge.com/article/DB047SECT10.

Haan, N., Smith, M. B., & Block, J. (1968). Moral reasoning of young adults: Political-social behavior, family background, and personality correlates. *Journal of Personality and Social Psychology, 10*(3), 183-201.

Harris, S., Mussen, P., & Rutherford, E. (1976). Maturity of moral judgment. *The Journal of Genetic Psychology: Research and Theory on Human Development*, *128*(1), 123-135.

Hatcher, J. A. (2008). *The public role of professionals: Developing and evaluating the civic- minded professional scale*. (Doctoral dissertation). Retrieved from Pro Quest Dissertation and Theses. (AAT 3331248).

Homiak, M. (2011). Moral Character. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy,* Stanford, CA: The Metaphysics Research Lab. Retrieved from http://plato.stanford.edu/archives/spr2011/entries/moral-character/.

Hoose, B. (1998). Charity. In *Routledge Encyclopedia of Philosophy Online*. London: Routledge. Retrieved from http://www.rep.routledge.com/article/L010SECT2.

Hu, L., & Bentler, P., M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

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

Jöreskog, K. G., & Sörbom, D. G. (2006). LISREL 8.80 [Computer software]. Lincolnwood, IL: Scientific Software.

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

King, P. M., & Mayhew, M. J. (2002). Moral judgment development in higher education: Insights from the defining issues test. *Journal of Moral Education, 31*(3), 247-270.

Keller, D. R. (2010). An introduction to ethics for teaching. *Teaching Ethics, 11*(1), 1-54.

Kline, R. B. (2011).*Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Kohlberg, L. (1958). *The development of modes of thinking and choices in years 10 to 16*. (Unpublished doctoral dissertation). University of Chicago, Illinois.

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.

Kohlberg, L. (1977). Implications of moral stages for adult education. *Religious Education, 72*(2), 183-201.

Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages*. San Francisco, CA: Harper & Row.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

Maitland, K. A., & Goldman, J. R. (1974). Moral judgment as a function of peer group interaction. *Journal of Personality and Social Psychology, 30*(5), 699-704.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). Differential Validity and Prediction of the SAT (College Board Research Report No. 2008-4). New York, NY: The College Board.

McDonald, R. P. (1999). *Test theory: A united treatment*. Mahwah, NJ: L. Erlbaum Associates.

McMahon, J. M., & Harvey, R. J. (2007).Psychometric properties of the Reidenbach-Robin multidimensional ethics scale. *Journal of Business Ethics, 72*(1), 27-39.

Mill, J. S. (2003). *On iberty* (Bromwich and Kateb ed.). London: Yale University Press. (Original work published in 1869).

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3). 548-570.

Page, R., & Bode, J. (1980). Comparison of measures of moral reasoning and development of a new objective measure. *Educational and Psychological Measurement, 40,* 317-329.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students*. San Francisco, CA: Jossey-Bass.

Patterson, B. F., & Mattern, K. D. (2013). Validity of the SAT for predicting first-year grades: 2010 SAT validity sample (College Board Research Report No. 2013-2). New York, NY: The College Board.

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

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods, 9*(1), 99-112.

Reidenbach, R. E., & Robin, D.P. (1988). Some initial steps toward improving the measurement of ethical evaluations of marketing activities. *Journal of Business Ethics, 7*, 871-879.

Reidenbach , R. E., & Robin, D.P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of Business Ethics, 9*, 639-653.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*, 129-140.

Rest, J. R. (1979b). *Development in judging moral issues.* Minneapolis, MN: University of Minnesota Press.

Rest, J. R., Cooper, D., Coder, R., Masanz, J., & Anderson, D. (1974). Judging the important issues in moral dilemmas: An objective measure of development. *Developmental Psychology, 10*(4), 491-501.

Rest, J. R., & Thoma, S. (1985). Relationship of moral judgment development to formal education, *Developmental Psychology, 21*, 709–714.

Rest, J. R., Deemer, D., Barnett, R., Spickelmier, J., & Volker, J. (1986). Life experiences and developmental pathways. In J. Rest (Ed.), *Moral development: Advances in research and theory* (p. 28-58). New York, NY: Praeger.

Rest, J. R., Narvaez, D., Bebeau, M., & Thoma, S. (1999). A neo-kohlbergian approach: The DIT and schema theory. *Educational Psychology Review, 11*(4), 291-324.

Rindskopf, D. (1984). Structural equation models: Empirical identification, heywood cases and related problems. *Sociological Methods and Research, 13*, 109-119.

Rubin, K. H., & Schneider, F. W. (1973). The relationship between moral judgment, egocentrism, and altruistic behavior. *Child Development, 44*(3), 661-665

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors on covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis*. Thousand Oaks, CA: Sage.

Schwartz, S. H., Feldman, K. A., Brown, M. E., & Heingartner, A. (1969). Some personality correlates of conduct in two situations of moral conflict. *Journal of Personality, 37*(1), 41-57.

Siegler, R. S. (1997). Concepts and methods for studying cognitive change. In Amsel, E., & Renninger, K. A. (Eds.), *Change and development: Issues of theory, method, and application.* Mahwah, NJ: Lawrence Erlbaum Associates.

Skorupski, J. (2005). Mill, John Stuart. In *Routledge Encyclopedia of Philosophy Online*. London: Routledge. Retrieved from http://www.rep.routledge.com/article/DC054.

Smart, J. J. C. & Williams, B. (1973). *Utilitarianism: For and against*. United Kingdom: Cambridge University Press.

Society for ethics across the curriculum. (2000). Retrieved August 2, 2013, from http://www.rit.edu/cla/ethics/seac/

Statement of purpose for the center for the study of ethics. (2013). Retrieved August 2, 2013, from http://www.uvu.edu/ethics/about/purpose.html

Statistical Definitions. (2013). Retrieved August 2, 2013, from http://research.collegeboard.org/definitions.

Steiger, J. H. (1980). Tests of comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245 - 251.

Steiger, J. H. (1990). Structural model evaluation and modification: An internal estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180.

Steiger, J. H., & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers, 4*, 581–582.

The Essential Learning Outcomes. (2013). Association of American Colleges and Universities. Retrieved from http://www.aacu.org/leap/documents/EssentialOutcomes_Chart.pdf

The Madison Collaborative: Ethical Reasoning in Action. (2013). Quality Enhancement Plan for the Southern Association of Colleges and Schools Commission on Colleges. Retrieved from http://www.jmu.edu/files/qep-proposal.pdf.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395-1415

Tong, R. (1998). Feminist ethics. In *Routledge Encyclopedia of Philosophy Online*. London: Routledge. Retrieved from http://www.rep.routledge.com/article/L026SECT1.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement and Practice, 10*(1), 37-45.

Treviño, L. K., & Nelson, K. A. (2011). *Managing business ethics: Straight talk about how to do it right* (5th ed.). New York, NY: John Wiley and Sons.

Welcome to the center for practical and professional ethics. (2006). Retrieved August 1, 2013, from http://www.csus.edu/cppe/index.html

Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist, 34*(5), 719 – 751.

Yost, A. B., & Finney, S. J. (In press). Building the nomological network of trait reactance using a multi-faceted model assessment approach. *Personality and Individual Differences.*

Yu, C., & Muthén, B (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Table 1
Factor loadings for one-factor solution

Item	Factor 1 Loading	Item	Factor 1 Loading	Item	Factor 1 Loading
1	0.375	21	0.651	41	0.575
2	0.300	22	0.162	42	0.625
3	0.430	23	0.354	43	0.429
4	0.487	24	0.401	**44**	**0.064**
5	0.393	25	0.548	45	0.613
6	0.452	26	0.255	46	0.588
7	0.284	27	0.803	47	0.254
8	0.549	28	0.380	48	0.567
9	0.467	29	0.589	49	0.276
10	0.306	30	0.577	50	0.551
11	0.366	31	0.468	51	0.228
12	0.433	32	0.627	52	0.240
13	0.871	33	0.480	53	0.355
14	0.425	34	0.658	54	0.372
15	0.375	35	0.683	55	0.561
16	0.736	36	0.520	56	0.221
17	0.499	37	0.405	57	0.606
18	0.359	38	0.689	58	0.887
19	0.606	39	0.516	59	0.644
20	0.601	40	0.449	60	0.397

**Note.* The bolded value represent the only item (#44) that did not have a significant loading onto the factor at α=.05.

Table 2
Factor loadings for 3-factor solution

Item	Factor 1 Loading	Factor 2 Loading	Factor 3 Loading	Item	Factor 1 Loading	Factor 2 Loading	Factor 3 Loading
1	0.229	0.149	0.223	31	0.097	0.488	0.114
2	**-0.125**	**0.594**	**-0.108**	32	0.586	0.051	0.315
3	0.289	0.191	0.112	**33**	**0.178**	**0.452**	**-0.235**
4	**0.616**	**-0.089**	**-0.218**	**34**	**0.646**	**-0.006**	**0.078**
5	-0.030	0.564	-0.083	35	0.677	0.02	0.151
6	0.429	0.036	0.018	36	0.279	0.313	0.061
7	0.097	0.235	-0.05	**37**	**0.156**	**0.302**	**0.166**
8	0.12	0.543	0.086	38	0.625	0.139	-0.076
9	**0.328**	**0.154**	**0.298**	**39**	**0.076**	**0.585**	**0.016**
10	0.301	0.02	0.089	40	0.098	0.402	0.244
11	0.283	0.101	0.088	41	0.434	0.188	0.034
12	**0.404**	**0.044**	**0.023**	42	0.644	-0.009	-0.067
13	0.928	-0.041	-0.158	43	0.266	0.172	0.150
14	**0.354**	**0.043**	**0.236**	44	0.102	-0.162	0.358
15	**0.269**	**0.164**	**-0.032**	45	0.459	0.166	0.205
16	**0.781**	**-0.051**	**0.028**	46	0.593	-0.021	0.148
17	0.173	0.457	-0.163	47	-0.074	0.486	-0.080
18	0.016	0.371	0.209	48	0.583	-0.022	-0.010
19	0.341	0.287	0.305	49	0.158	0.183	-0.170
20	**0.496**	**0.131**	**0.050**	50	0.551	0.015	0.126
21	0.524	0.190	-0.020	**51**	**0.079**	**0.189**	**-0.027**
22	**-0.171**	**0.318**	**0.393**	52	0.222	0.027	0.055
23	**0.368**	**-0.065**	**0.196**	53	0.267	0.239	-0.413
24	0.331	0.106	0.087	54	0.314	0.116	-0.183
25	**-0.001**	**0.662**	**0.102**	55	0.358	0.298	-0.048
26	0.168	0.085	0.059	**56**	**0.124**	**0.144**	**-0.037**
27	**0.782**	**0.009**	**0.138**	57	0.645	0	-0.100
28	**0.068**	**0.439**	**-0.155**	**58**	**0.826**	**0.160**	**-0.115**
29	0.549	0.082	-0.021	59	0.762	-0.151	0.020
30	0.571	0.014	-0.054	60	0.416	-0.010	0.054

**Note.* The bolded values represent the 20 items written to align with the Personal "area of application." These items don't appear to share meaningful, common variance.

Table 3
Factor loadings for 8-factor solution

Item	Factor 1 Loading	Factor 2 Loading	Factor 3 Loading	Factor 4 Loading	Factor 5 Loading	Factor 6 Loading	Factor 7 Loading	Factor 8 Loading
1	-0.102	0.340	-0.033	0.143	0.034	0.075	0.067	0.102
2	-0.021	-0.235	0.102	0.123	0.607	-0.137	0.017	0.228
3	0.132	-0.009	0.097	0.346	0.180	-0.032	-0.115	0.042
4	0.644	0.07	-0.003	-0.027	0.049	0.002	0.061	-0.032
5	-0.088	0.049	0.576	-0.066	0.117	0.146	0.042	0.03
6	0.314	0.042	0.011	0.221	0.093	-0.016	-0.045	0.043
7	0.040	-0.042	0.013	0.146	0.231	-0.023	0.078	0.026
8	0.041	0.265	0.098	-0.06	0.496	0.029	0.022	-0.024
9	0.185	0.582	0.122	-0.036	-0.007	-0.053	-0.093	0.043
10	0.181	0.06	0.103	0.134	0.06	0.085	-0.117	-0.189
11	0.048	0.062	0.136	0.346	-0.038	0.025	0.019	-0.049
12	**0.131**	**0.072**	**0.086**	**0.121**	**-0.08**	**0.179**	**-0.012**	**0.325**
13	0.701	0.09	0.036	0.074	0.067	0.197	-0.024	0.233
14	0.215	0.303	0.181	0.197	-0.123	-0.071	-0.157	0.072
15	0.230	0.096	0.167	0.110	0.085	-0.021	0.03	-0.12
16	0.320	0.045	0.05	0.372	0.009	0.309	0.002	-0.149
17	-0.057	-0.006	0.360	-0.051	0.175	0.325	0.151	0.052
18	-0.314	0.453	0.081	-0.047	0.123	0.143	0.144	0.048
19	0.049	0.580	0.092	0.072	0.088	0.008	0.02	0.063
20	0.288	0.172	0.045	0.313	0.115	-0.020	0.079	-0.13
21	**0.195**	**0.350**	**0.009**	**0.004**	**0.106**	**0.207**	**0.251**	**-0.031**
22	-0.359	0.188	0.347	0.310	-0.084	-0.069	-0.208	0.117
23	-0.086	0.212	-0.290	0.274	0.058	0.164	0.099	0.016
24	**-0.021**	**0.052**	**0.01**	**0.230**	**0.012**	**0.168**	**0**	**0.340**
25	-0.169	0.167	0.135	0.034	0.566	0.099	0.032	-0.034
26	0.049	0.03	0.132	0.037	0.063	0.142	-0.103	-0.076
27	0.341	0.350	-0.056	0.206	0.015	0.176	0.012	0.21
28	0.168	0.001	0.541	0.034	0.049	-0.056	0.075	-0.026
29	**0.249**	**0.136**	**-0.063**	**0.250**	**0.102**	**0.094**	**0.160***	**0.038**
30	0.360	0.028	0.16	0.167	-0.015	0.172	-0.027	-0.01
31	-0.062	0.179	0.007	0.185	0.426	-0.060	0.078	0.086
32	0.093	0.470	-0.067	0.168	0.049	0.222	-0.025	-0.008
33	0.158	-0.131	0.171	0.208	0.311	-0.056	0.223	0.095
34	0.219	-0.075	-0.07	0.515	0.09	0.167	-0.036	0.119
35	0.244	0.201	0.012	0.350	0.058	0.200*	-0.009	-0.138
36	0.232	0.206	0.339	0.042	0.109	0.007	-0.09	-0.017
37	-0.121	0.152	0.147	0.469	0.081	-0.093	0.092	-0.055
38	0.369	0.104	0.033	0.202	0.195	0.15	0.117	-0.138

39	0.08	0.004	-0.016	-0.059	0.820	0.045	-0.093	-0.059
40	-0.132	0.290	0.017	0.180	0.303	-0.023	0.008	0.067
41	0.161	0.074	-0.014	0.298	0.151	0.037	0.065	0.243
42	**0.421**	**0.107**	**0.014**	**0.142**	**0.005**	**0.114**	**0.058**	**0.166**
43	-0.031	0.085	0.167	0.512	-0.041	-0.044	0.023	-0.026
44	0.056	-0.028	0.012	0.041	0.111	0.129	-0.750	-0.003
45	0.007	0.163	-0.085	0.412	0.184	0.12	0.024	0.077
46	-0.048	0.048	-0.094	0.283	0.121	0.511	-0.043	-0.102
47	-0.026	0.022	0.435	-0.015	0.088	-0.032	0.059	0.201
48	0.226	0.006	-0.09	0.221	0.089	0.232	0.03	0.133
49	0.334	0.069	0.239	-0.033	0.016	-0.151	0.115	-0.032
50	0.094	0.109	0.094	0.258	0.001	0.333	-0.062	-0.103
51	-0.06	-0.04	0.253	0.093	0.019	0.161	0.017	-0.118
52	-0.123	0.045	0.015	0.087	0	0.307	0.037	-0.044
53	0.063	-0.199	0.167	0.143	-0.011	0.210	0.490	0.074
54	0.306	0.013	0.230	-0.003	0.034	0.094	0.109	-0.237
55	0.031	0.139	0.09	0.12	0.159	0.209	0.200	0.068
56	**0.072**	**0.04**	**0.028**	**-0.11**	**-0.003**	**0.036**	**0.03**	**0.676**
57	0.063	-0.037	0.173	0.161	-0.097	0.594	0.111	-0.043
58	**0.213**	**0.150**	**0.028**	**-0.088**	**0.234**	**0.681**	**0.152**	**0.025**
59	0.04	0.053	-0.047	-0.003	-0.04	0.781	-0.061	0.122
60	-0.069	-0.059	0.117	-0.011	0.017	0.598	-0.185	0.109

**Note.* The bolded values represent the 7 items written to align with the Empathy "Key Question." These items don't appear to share meaningful, common variance.

Table 4
Factor intercorrelations for 8-factor EFA model

factor	1	2	3	4	5	6	7	8
1	1							
2	0.316*	1						
3	0.317*	0.308*	1					
4	0.086	0.292*	0.209*	1				
5	0.365*	0.280*	0.365*	0.219*	1			
6	0.504*	0.325*	0.378*	0.165*	0.388*	1		
7	0.076	0.092	0.003	0.107	-0.011	0.076	1	
8	0.071	0.125*	0.054	0.112*	0.096*	0.158	-0.025	1

**Note.* Asterisk denotes significant correlations.

Table 5
Reliability information for all administrations of the ERIT

	Fall 2012 Administration (Freshmen)		Spring 2013 Administration (Sophomores)	Fall 2013 Administration (Freshmen)	
	TERA	TERB	ERIT-0	ERIT-1 (all 50 items)	ERIT-1 (without 8 testlet items)
# Items	48	48	60	50	42
N	446	432	793	809	829
Mean	61.72%	62.17%	66.05%	69.00%	68.19%
SD	12.35%	10.92%	14.60%	13.07%	13.65%
Coefficient alpha (KR-20)	.764	.691	.872	.809	.787
Spearman-Brown Prophecy Formula predicted alpha (60 item)	.802	.737	---	---	---
Spearman-Brown Prophecy Formula predicted alpha (50 item)	.771	.700	---	---	---
McDonald's Omega	---	---	---	---	.792

**Note*. The TERA and TERB were administered during Fall 2012, the ERIT-0 was administered during Spring 2013, and the ERIT-1 was administered during Fall 2013.

Table 6
Data collection design

	Fall 2012	Fall 2013
Research Question		
1		ERIT-1
2		ERIT-1 & SAT-CR
3		ERIT-1 & ERRT
4	TERA & TERB	ERIT-1

**Note*. ERIT-1 refers to the Ethical Reasoning Identification Test version 1adminstered during fall 2013, SAT-CR refers to the SAT Critical Reading test (formerly verbal proficiency), ERRT refers to the Ethical Reasoning Recall Test, TERA refers to the 48-item version A of the ERIT administered during fall 2012, and TERB refers to the 48-item version B of the ERIT administered during fall 2012.

Table 7
Demographic information for students completing the TERA, TERB, ERIT-0, ERIT-1, and ERRT

	TERA	TERB	ERIT-0	ERIT-1 (all 50-items)	ERIT-1 (42-items)	ERRT
N	446	432	793	809	829	140
Age	18.44 (0.37)	18.43 (0.36)	20.18 (1.05)	18.45 (0.37)	18.44 (0.37)	18.39 (0.34)
SAT verbal	568.47 (66.22)	571.33 (67.464)	572.11 (70.89)	571.83 (67.40)	571.28 (67.43)	568.55 (65.26)
SAT math	578.28 (69.10)	581.55 (64.79)	581.09 (67.70)	577.72 (64.49)	577.28 (64.91)	580.61 (60.10)
Female	60.31%	57.64%	58.8%	60.8%	61.2%	66.4%
Caucasian	86.3%	87.7%	87.1%	87.5%	87.5%	90.0%
Asian	6.3%	6.7%	5.6%	5.7%	6.0%	6.4%
African American	5.8%	4.2%	5.6%	5.1%	4.9%	2.1%
Hispanic	4.9%	6.0%	3.9%	4.5%	4.5%	5.0%
Native American Indian	1.1%	0.7%	1.3%	1.5%	1.4%	2.1%
Pacific Islander	0.2%	1.2%	0.6%	0.4%	0.4%	0%

**Note.* Means are presented with standard deviations in parentheses. Listwise deletion was used for all tests.

Table 8.
Difficulty and alpha if deleted values for ERIT-1 items

KQ	Item	Difficulty	Std Dev	Alpha if deleted	KQ	Item	Difficulty	Std Dev	Alpha if deleted
Rights	39	0.171	0.377	0.787	Responsibilities	11	0.660	0.474	0.787
Rights	42	0.287	0.453	0.792	Outcomes	16	0.710	0.454	0.785
Rights	2	0.422	0.494	0.779	Fairness	3	0.712	0.453	0.784
Authority	26	0.475	0.500	0.781	Rights	22	0.731	0.444	0.783
Liberty	14	0.497	0.500	0.780	Empathy	23	0.756	0.430	0.782
Outcomes	10	0.505	0.500	0.784	Authority	15	0.767	0.423	0.780
Authority	6	0.520	0.500	0.783	Character	40	0.796	0.403	0.782
Liberty	5	0.520	0.500	0.779	Empathy	19	0.808	0.394	0.786
Responsibilities	30	0.520	0.500	0.782	Outcomes	1	0.842	0.365	0.781
Responsibilities	36	0.538	0.499	0.781	Responsibilities	37	0.855	0.352	0.782
Responsibilities	29	0.544	0.498	0.783	Authority	25	0.862	0.345	0.780
Liberty	24	0.545	0.498	0.783	Fairness	38	0.873	0.333	0.783
Liberty	33	0.552	0.498	0.782	Character	34	0.888	0.316	0.781
Character	41	0.559	0.497	0.787	Character	27	0.902	0.297	0.783
Outcomes	7	0.569	0.495	0.791	Empathy	17	0.905	0.294	0.782
Liberty	32	0.592	0.492	0.781	Outcomes	28	0.925	0.263	0.783
Liberty	20	0.609	0.488	0.776	Character	9	0.928	0.259	0.785
Responsibilities	8	0.609	0.488	0.788	Empathy	35	0.929	0.257	0.781
Rights	4	0.626	0.484	0.783	Fairness	31	0.940	0.238	0.783
Rights	13	0.626	0.484	0.782	Fairness	12	0.953	0.212	0.786
Fairness	18	0.651	0.477	0.783	Character	21	0.959	0.198	0.784

Table 9
Standardized factor pattern coefficients and variance explained for one-factor model

Item	Factor Pattern Coefficient	Std Error	t-value	R^2	Item	Factor Pattern Coefficient	Std Error	t-value	R^2
1	0.49	0.054	9.20*	0.240	22	0.37	0.065	5.73*	0.140
2	0.51	0.053	9.63*	0.260	23	0.44	0.055	8.04*	0.190
3	0.35	0.059	6.00*	0.120	24	0.39	0.053	7.27*	0.150
4	0.36	0.056	6.44*	0.130	25	0.63	0.059	10.67*	0.400
5	0.50	0.052	9.63*	0.250	26	0.46	0.048	9.67*	0.210
6	0.38	0.054	7.04*	0.140	27	0.49	0.070	6.96*	0.240
7	0.09	0.062	1.39	0.007	28	0.60	0.071	8.50*	0.360
8	0.18	0.062	2.86*	0.032	29	0.36	0.059	6.11*	0.130
9	0.42	0.090	4.73*	0.180	30	0.38	0.061	6.18*	0.140
10	0.32	0.061	5.24*	0.100	31	0.60	0.088	6.86*	0.370
11	0.22	0.062	3.57*	0.050	32	0.46	0.054	8.68*	0.220
12	0.40	0.110	3.55*	0.160	33	0.43	0.055	7.72*	0.180
13	0.42	0.053	7.79*	0.170	34	0.59	0.070	8.37*	0.350
14	0.50	0.053	9.43*	0.250	35	0.73	0.071	10.29*	0.530
15	0.53	0.054	9.82*	0.290	36	0.42	0.046	9.06*	0.180
16	0.27	0.063	4.37*	0.075	37	0.49	0.065	7.51*	0.240
17	0.60	0.066	9.08*	0.360	38	0.45	0.076	5.93*	0.200
18	0.36	0.056	6.50*	0.130	39	0.25	0.075	3.41*	0.065
19	0.30	0.067	4.41*	0.088	40	0.44	0.063	7.05*	0.190
20	0.63	0.043	14.89*	0.400	41	0.21	0.062	3.40*	0.044
21	0.62	0.100	6.18*	0.390	42	0.05	0.065	0.77	0.003

**Note.* * denotes path that is statistically significant at $\alpha = .01$. The correlation matrix was analyzed, so the parameter estimates were standardized.

Table 10.
Correlation residuals greater than |.2| for one-factor model compared to bifactor model

Item		Common Key Question	Common Application Area	Correlation residual for one-factor model	Correlation residual for bifactor model
1	21	---	---	-0.275	-0.280
2	13	Rights	---	0.230	0.220
4	28	---	---	-0.224	-0.230
7	16	Outcomes	---	0.215	0.210
9	38	---	---	-0.216	-0.220
11	19	---	---	-0.258	-0.260
11	30	Responsibilities	Personal	0.233	0.230
13	14	---	---	---	0.220
13	31	---	Professional	-0.216	-0.220
14	31	---	---	-0.253	-0.210
15	25	Authority	---	0.226	0.231
28	36	---	---	---	-0.210
30	36	Responsibilities	---	0.234	0.230
31	37	---	Professional	0.235	0.220

**Note.* The residuals between items 13 and 14 and between items 28 and 36 were less than |.2| for the one-factor model.

Table 11
Modification indices for one-factor model

Item 1	Item 2	Decrease in Chi-Square	New Estimate	Correlation Residual	Common KQ	Common Application Area	Theoretically plausible?
25	**15**	**68.7**	**0.88**	**0.226**	**Authority**		**yes**
15	6	21.7	0.35	0.182	Authority		yes
26	15	9.8	0.23	0.226	Authority		yes
40	34	17.4	0.43	0.177	Character	Professional	yes
23	17	11.8	0.32	0.173	Empathy	Civic	yes
35	17	141.4	4.84	0.143	Empathy		yes
38	18	15.1	0.32	0.196	Fairness		yes
32	20	39.8	0.56	0.164	Liberty	Personal	yes
20	14	596.8	5.86	0.200	Liberty		yes
33	20	90.9	1.03	0.193	Liberty		yes
20	5	84.6	0.95	0.180	Liberty		yes
14	5	41.1	0.51	0.182	Liberty		yes
32	5	31.9	0.46	0.169	Liberty		yes
7	1	7.9	-0.21	-0.161	Outcomes	Civic	yes
16	10	9.8	0.20	0.150	Outcomes	Personal	yes
16	**7**	**16.2**	**0.24**	**0.215**	**Outcomes**		**yes**
16	1	13.9	0.29	0.182	Outcomes		yes
30	**11**	**26.5**	**0.32**	**0.233**	**Responsibilities**	**Personal**	**yes**
36	**30**	**48.6**	**0.49**	**0.234**	**Responsibilities**		**yes**
13	**2**	**52.5**	**0.54**	**0.230**	**Rights**		**yes**
13	4	11.5	0.22	0.145	Rights		yes
27	11	7.9	-0.24	-0.199	---	Personal	yes
37	**31**	**84.4**	**1.91**	**0.235**	**---**	**Professional**	**yes**
40	35	10.2	0.37	0.144	---	Professional	yes
37	5	9.1	-0.25	-0.165	---	Professional	yes
31	5	8.7	-0.32	-0.199	---	Professional	yes
31	**13**	**8.6**	**-0.31**	**-0.216**	**---**	**Professional**	**yes**
34	25	7.9	-0.29	-0.185	---	Professional	yes
14	13	27.8	0.38	0.180	---		no
23	3	17.9	0.31	0.193	---		no
19	11	17.8	-0.29	-0.258	---		no
36	26	15.4	0.26	0.154	---		no
37	20	14	-0.34	-0.187	---		no
31	14	13.6	-0.40	-0.253	---		no
37	18	12.8	0.29	0.169	---		no
5	4	12.3	0.25	0.135	---		no
24	22	11.7	0.24	0.155	---		no
6	1	11.6	0.26	0.160	---		no
35	26	10.9	0.33	0.182	---		no
28	4	10	-0.31	-0.244	---		no

10	4	9.7	-0.20	-0.144	---	no
28	27	9.6	0.36	0.176	---	no
22	15	9.5	-0.26	-0.152	---	no
34	17	9	0.38	0.128	---	no
29	27	8.7	-0.26	-0.186	---	no
20	15	8.4	-0.24	-0.124	---	no
31	20	8.3	-0.35	-0.166	---	no
26	14	8.2	-0.19	-0.122	---	no
36	28	7.9	-0.26	-0.199	---	no

Note. Bolded values indicate modification indices suggested by LISREL that also had correlation residuals greater than |.2| for one-factor model and the bifactor model (See Table 10).

Table 12
Factor pattern coefficients for one-factor model compared to bifactor model

one-factor				Bifactor					
Item	Pattern Coefficient for Ethical Reasoning Factor	Item	Pattern Coefficient for Ethical Reasoning Factor	Item	Pattern Coefficient for Ethical Reasoning Factor	Pattern Coefficient for Liberty* subfactor	Item	Pattern Coefficient for Ethical Reasoning Factor	Pattern Coefficient for Liberty* subfactor
1	0.49	22	0.37	1	0.50	- -	22	0.38	- -
2	0.51	23	0.44	2	0.52	- -	23	0.45	- -
3	0.35	24	0.39	3	0.36	- -	24	0.34	0.27
4	0.36	25	0.63	4	0.36	- -	25	0.65	- -
5	0.50	26	0.46	5	0.40	0.50	26	0.48	- -
6	0.38	27	0.49	6	0.39	- -	27	0.50	- -
7	0.09	28	0.60	7	0.09	- -	28	0.61	- -
8	0.18	29	0.36	8	0.18	- -	29	0.37	- -
9	0.42	30	0.38	9	0.43	- -	30	0.39	- -
10	0.32	31	0.60	10	0.32	- -	31	0.62	- -
11	0.22	32	0.46	11	0.23	- -	32	0.39	0.37
12	0.40	33	0.43	12	0.41	- -	33	0.36	0.34
13	0.42	34	0.59	13	0.42	- -	34	0.60	- -
14	0.50	35	0.73	14	0.41	0.46	35	0.74	- -
15	0.53	36	0.42	15	0.55	- -	36	0.43	- -
16	0.27	37	0.49	16	0.28	- -	37	0.51	- -
17	0.60	38	0.45	17	0.61	- -	38	0.46	- -
18	0.36	39	0.25	18	0.37	- -	39	0.26	- -
19	0.30	40	0.44	19	0.30	- -	40	0.45	- -
20	0.63	41	0.21	20	0.52	0.64	41	0.22	- -
21	0.62	42	0.05	21	0.63	- -	42	0.05	- -

Table 13
Correlation residuals for Liberty items for one-factor model compared to bifactor model

Items		One-factor	Bifactor	Reduced?
20	32	0.16	0.11	Yes
20	5	0.18	0.03	Yes
20	24	0.07	0.04	Yes
20	14	0.20	0.01	Yes
20	33	0.15	0.06	Yes
32	5	0.17	0.05	Yes
32	24	0.11	0.05	Yes
32	14	0.02	0.09	No
32	33	0.02	0.05	No
5	24	0.08	0	Yes
5	14	0.18	0.03	Yes
5	33	0.08	0.03	Yes
24	14	0.11	0.03	Yes
24	33	0.03	0.02	Yes
14	33	0.09	0.01	Yes

Table 14

Comparison of difficulty values for students that experienced an ER intervention and students that did not

ERIT-1 Item #	Difficulty for Intervention Group	TERA/TERB Item #	Difficulty for No Intervention Group	Intervention - No intervention
10	0.505	24	0.614	-0.109
11	0.660	27	0.765	-0.106
36	0.538	38	0.597	-0.059
8	0.609	17	0.667	-0.058
29	0.544	25	0.590	-0.046
3	0.712	11	0.757	-0.045
1	0.842	5	0.874	-0.032
12	0.953	29	0.972	-0.019
28	0.925	24	0.942	-0.017
31	0.940	29	0.938	0.002
26	0.475	19	0.470	0.005
9	0.928	22	0.913	0.015
17	0.905	37	0.868	0.037
6	0.520	15	0.473	0.047
15	0.767	34	0.710	0.057
24	0.545	13	0.470	0.075
33	0.552	33	0.470	0.083
30	0.520	27	0.432	0.088
25	0.862	18	0.773	0.090
13	0.626	30	0.535	0.091
27	0.902	22	0.782	0.121
14	0.497	33	0.350	0.147
21	0.959	8	0.795	0.164
5	0.520	13	0.316	0.204
20	0.609	4	0.383	0.226
32	0.592	32	0.361	0.231

Note*. Values highlighted in purple are items from the TERA. Values highlighted in **gold are items from the **TERB**. The nine items highlighted in red are items that a higher percentage of "No Intervention" students responded correctly to. As item colors transition from dark red to lighter red to lighter green to **dark green**, a larger percentage of "Intervention Students" responded correctly to that item compared to the "No Intervention" students. For example, the items in **dark green** are the items that at least 20% more "Intervention Students" responded correctly to compared to "No Intervention" students.

Figure 1. Comprehensive ethical reasoning intervention plan

1= Light Exposure, 2 = Moderate Exposure, 3 = Heavy Exposure

Required Professional Development for Implementers		Core Module	Core Module	Core & Co-Curricular Modules	Core & Curricular Modules	Core & Curricular Modules
Indirect Interventions	*The One Book* and other communication	⇩	⇩	⇩	⇩	⇩
Direct Interventions	⇩	It's Complicated: Ethical Reasoning in Action	MC Freshman Course	Residence Life Scenarios	Gen Ed Course, Ethical Reasoning Infused	Course in Major, Ethical Reasoning Infused
SLO 1 Memorization		1	2	1	2	2
SLO 2 Identification Simple		1	3	2	2	2
SLO 3 Identification Complex		1	3		2	2
SLO 4 Application Generic		1	1	1	3	3
SLO 5 Application Personal			1		1	1
SLO 6 Importance	1	2	2	1	1	1
SLO 7 Confidence			1		1	1
% of students affected during career	99% of freshmen	99% of freshmen	99.9% of freshmen	Approx. 50% of freshmen & sophomores	Approx. 76% of all students*	Approx. 20% of all students
Intervention initiation (on some scale)	Summer 2013	Fall 2013	Fall 2014	Fall 2013	Fall 2013	Fall 2013

*10% of incoming freshmen and most transfer students will bypass General Education courses by obtaining Advanced Placement, International Baccalaureate, or transfer credit for these classes. Transfer students account for, on average, 16% of the JMU population.

Figure 2. Mapping of assessment tools to student learning outcomes

	SLO 1: Memorization of 8KQs	**SLO 2 & 3:** Identifying relationship of KQs to a decision or rationale	**SLO 4:** Applying KQs to a specific hypothetical situation or dilemma	**SLO 5:** Applying KQs to students' own personal, professional, or civic ethical cases	**SLO 6 & 7:** Attitudes toward ethical reasoning	**Data Collection:**
Ethical Reasoning Recall Test (Direct Measure)	X					Data collected on 100-200 randomly selected students on assessment days as beginning freshmen and again as sophomores/juniors. Repeated-Measures Design
Ethical Reasoning Identification Test (Direct measure)		X				Data collected on 500-1000 randomly selected students on assessment days as beginning freshmen and again as sophomores/juniors. Repeated-Measures Design
Ethical Reasoning Essay (Direct Measure)				X		Data collected on 100-200 randomly selected students on assessment days as beginning freshmen and again as sophomores/juniors. Repeated-Measures Design
Survey of Ethical Reasoning (Indirect Measure)	X	X	X	X	X	Data collected on 500-1000 randomly selected students on assessment days as beginning freshmen and again as sophomores/juniors. Repeated-Measures Design

Figure 3. Example of content map for ERIT-1

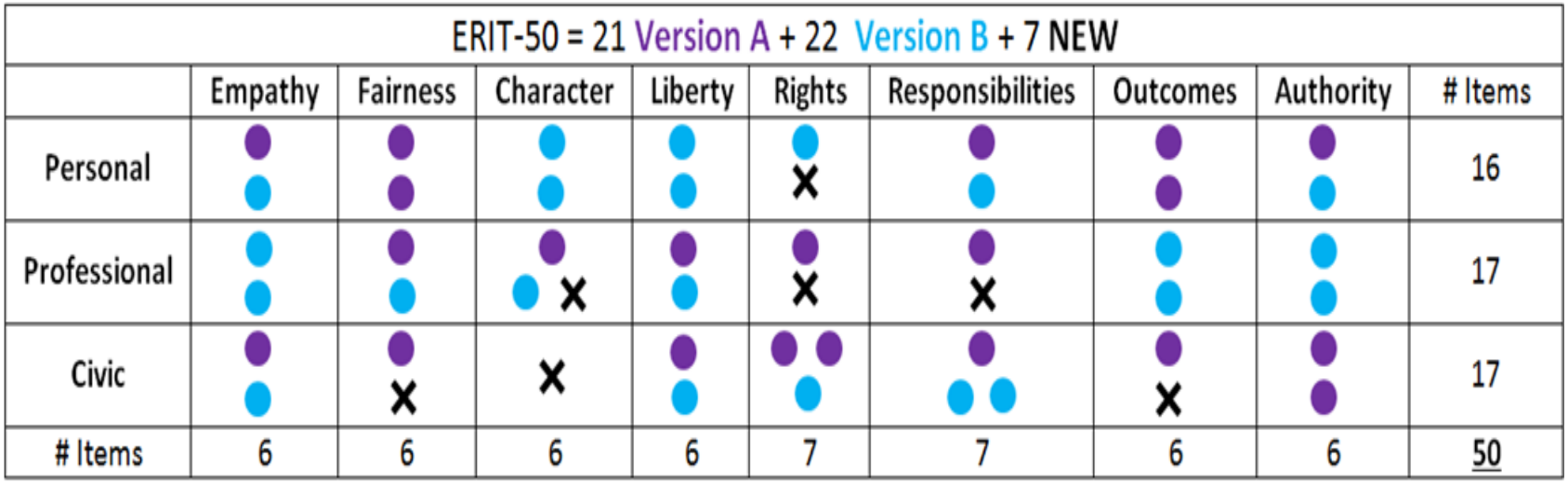


Figure 4. Scree plot of eigenvalues from factor analysis using tetrachoric correlation matrix

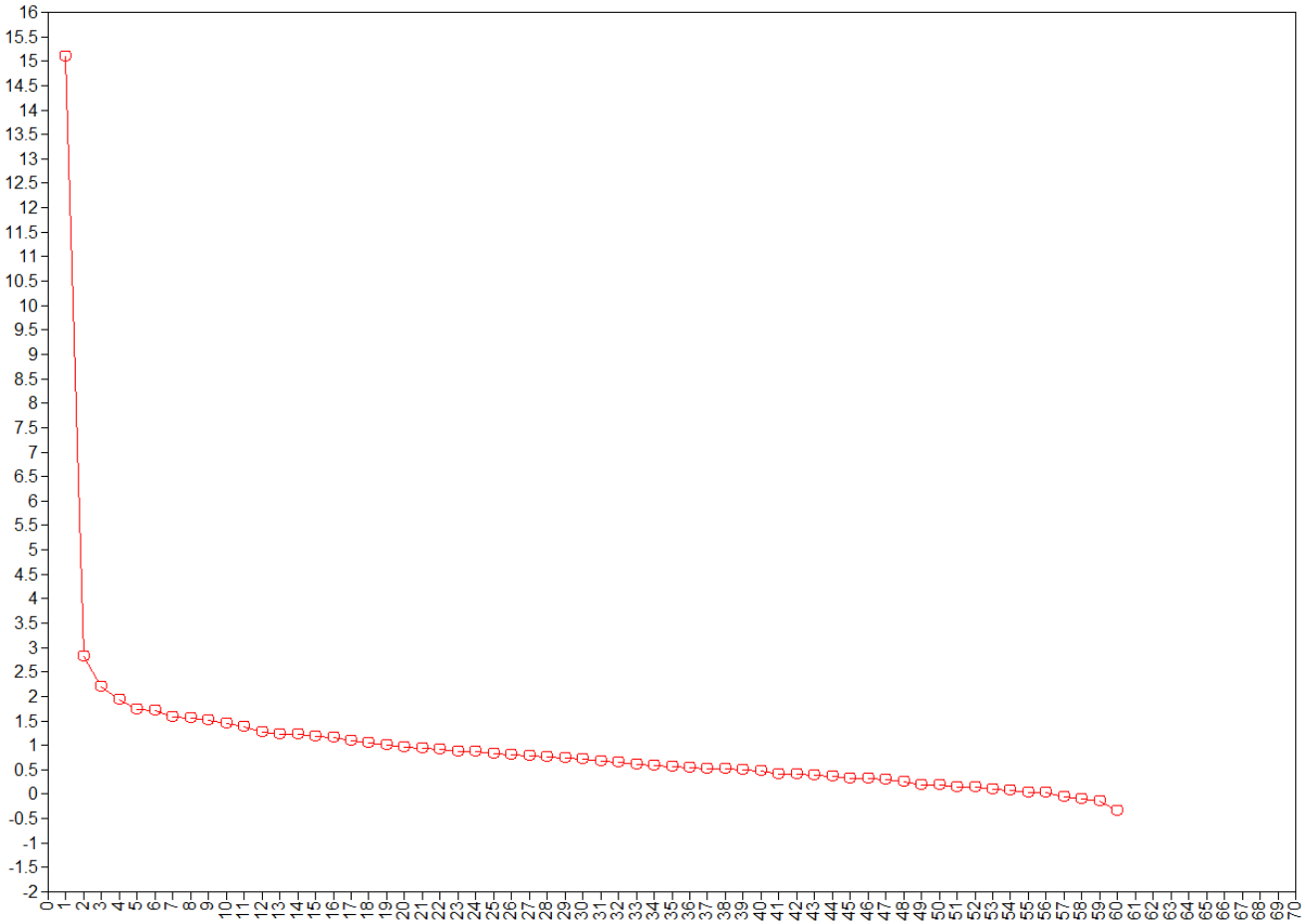


Figure 5. One-factor model

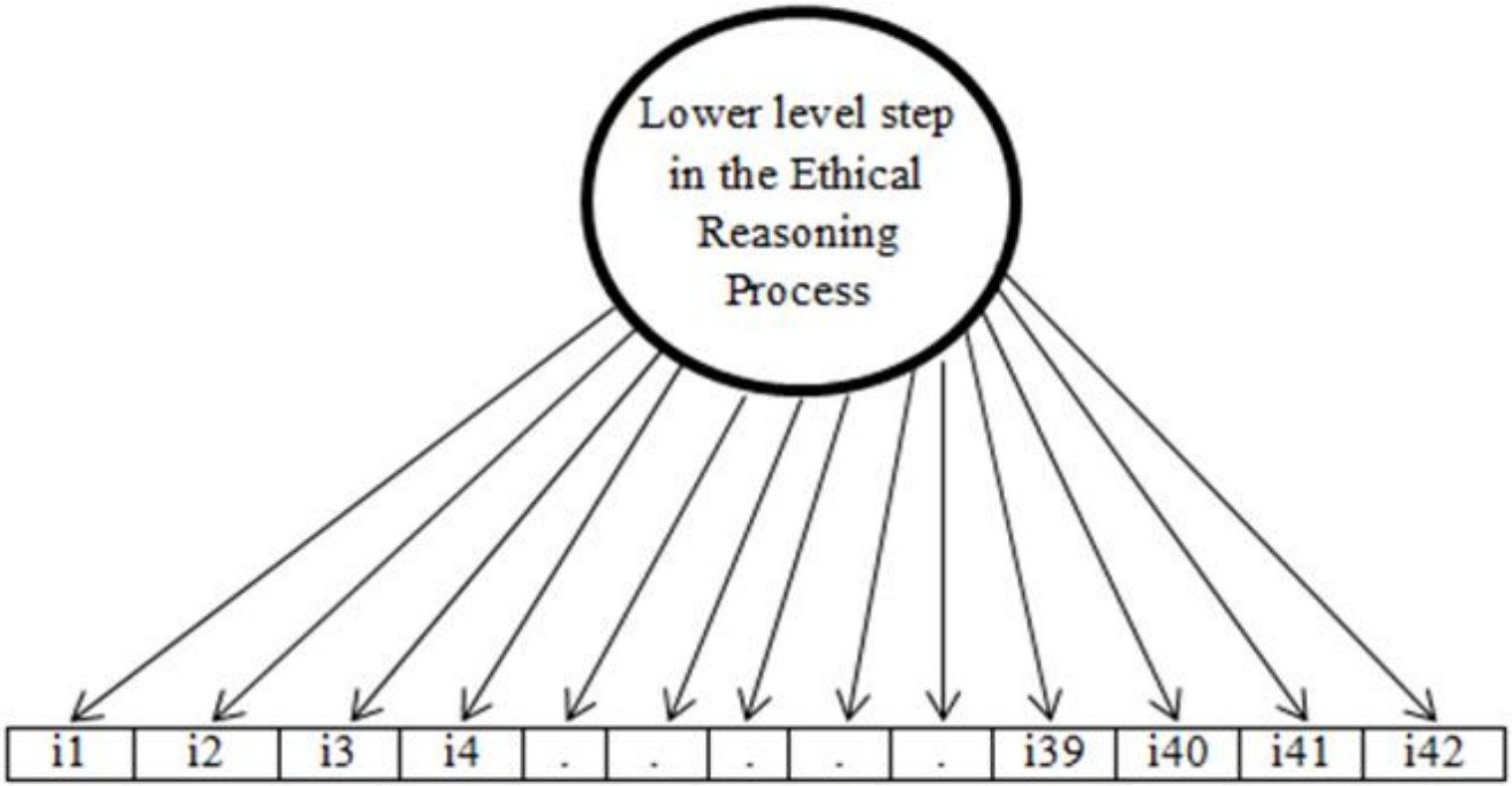


Figure 6. 3-factor model

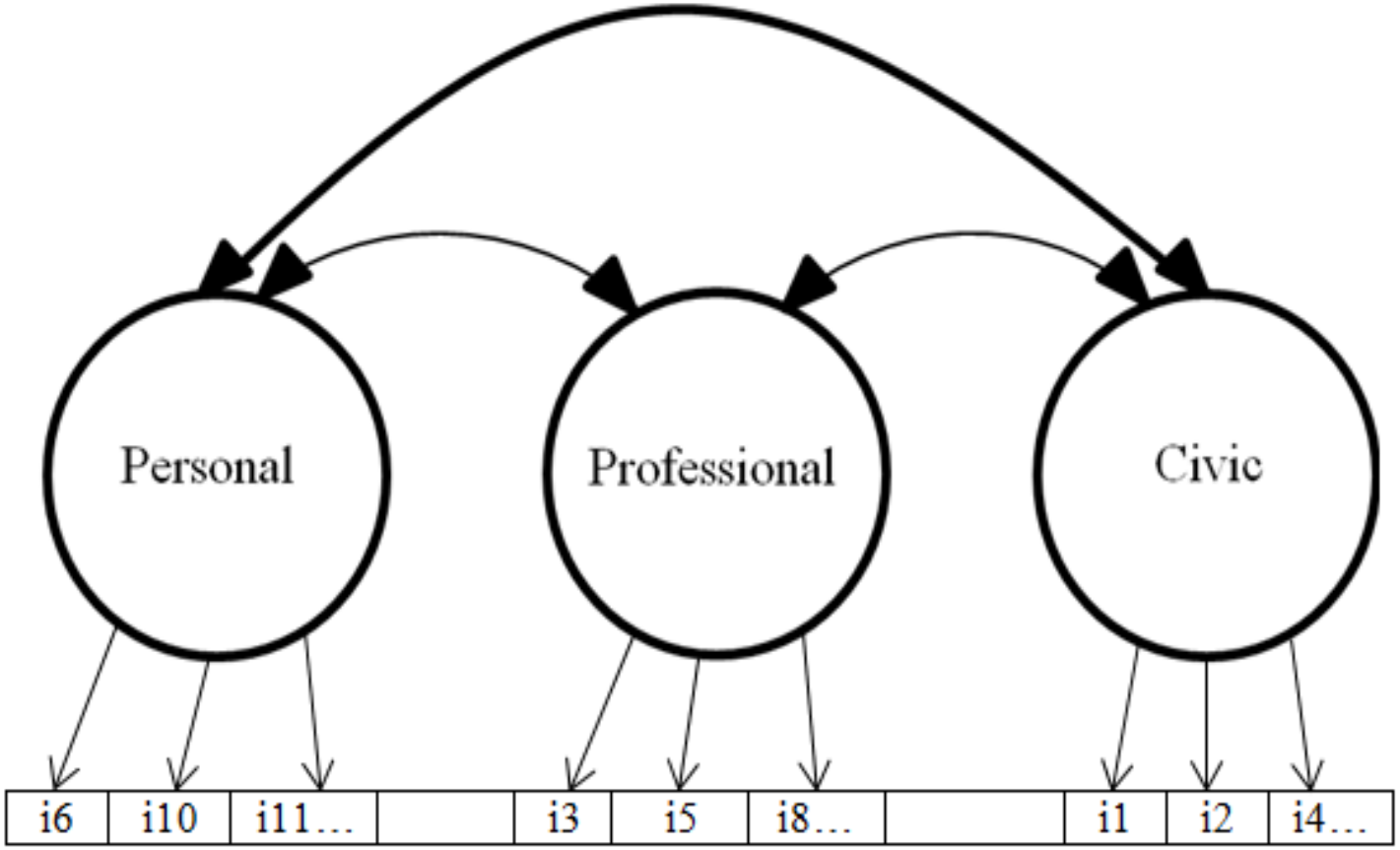
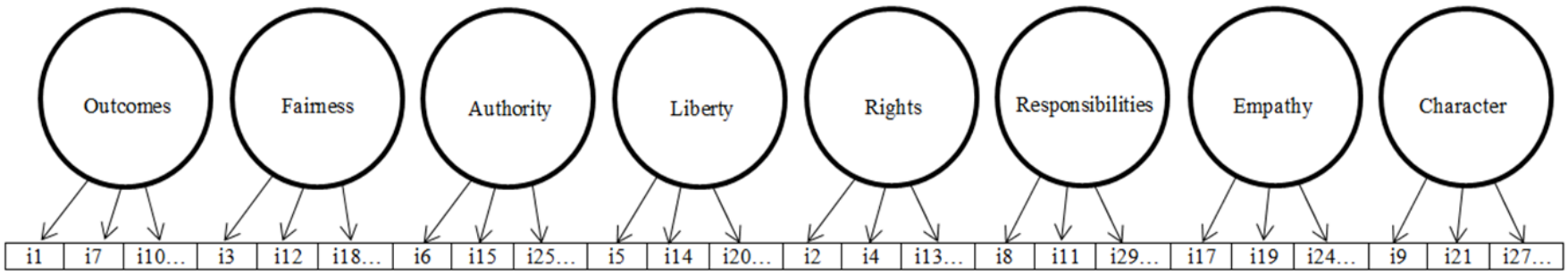
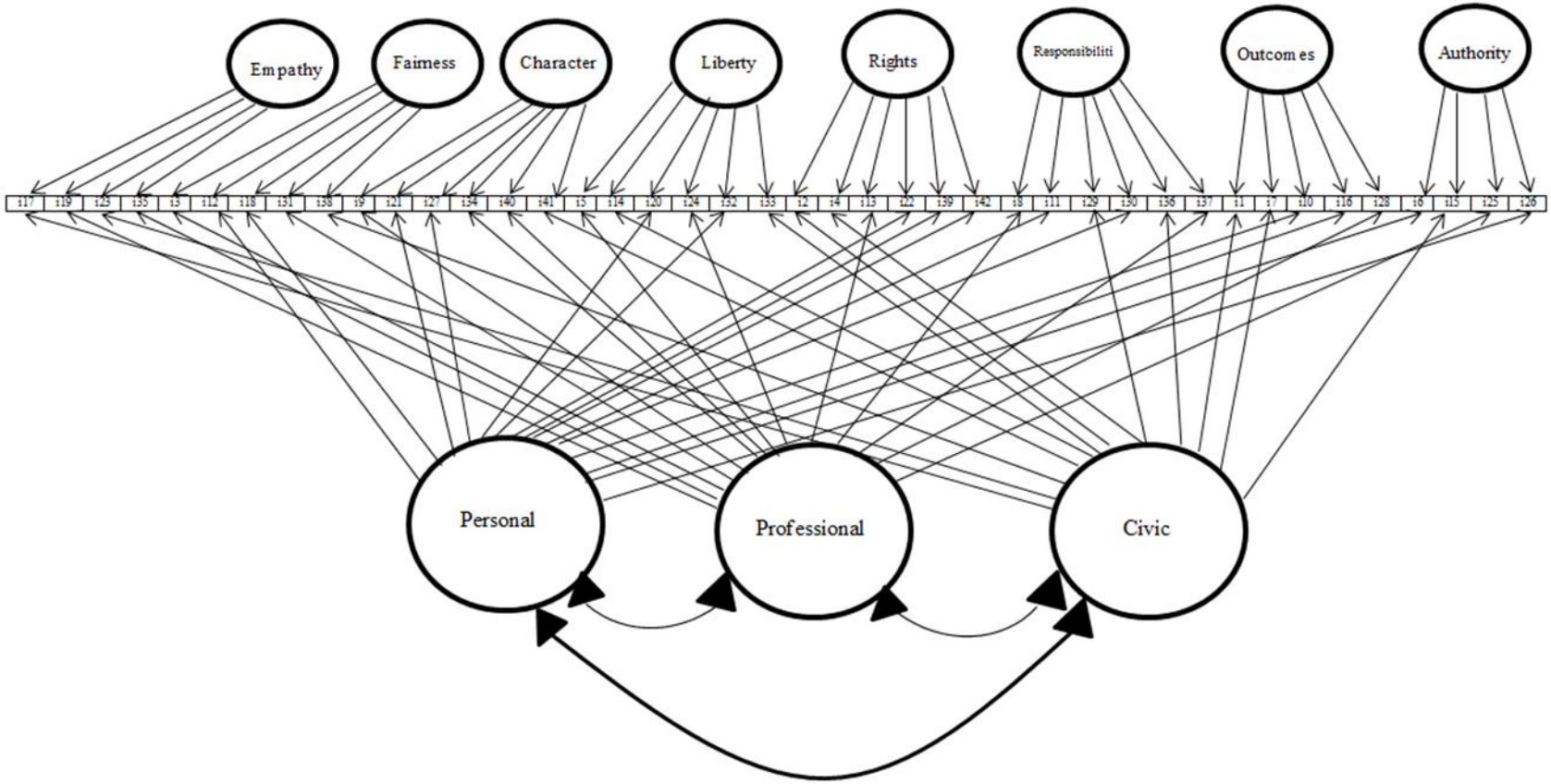


Figure 7. 8-factor model



Note. Latent variables were allowed to freely correlate. They are all intercorrelated to some extent. Intercorrelations between 8 KQ latent factors are not depicted

Figure 8. 3*8-factor hierarchical model



Note. The 8 KQ latent variables were allowed to freely correlate with each of the other KQ latent variables. Similarly, the 3 application area latent variables were allowed to freely correlate with each of the other application area latent variables; however, the 8 KQ latent variables were not allowed to correlate with any of the 3 application area latent variables. Intercorrelations between 8 KQ latent factors are not depicted in figure.

Figure 9. “Liberty*” bifactor model