Spring 2012

Demonstrating validity evidence of meta-assessment scores using generalizability theory

Chris D. Orem
James Madison University

Demonstrating Validity Evidence of Meta-Assessment Scores Using Generalizability Theory

Chris D. Orem

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2012

Acknowledgements

I would first like to acknowledge my dissertation committee: Dr. Donna Sundre, Dr. Christine DeMars, Dr. Herb Amato, and Dr. Keston Fulcher. They committed their time and energy to ensuring my success with this project, and I am grateful to them for their dedication and guidance. In particular, I want to thank my advisor and committee chair, Dr. Keston Fulcher, for his continued support throughout this process. You pushed me to be a better student and scholar and for that I am grateful. Christine, thank you for your patience and skill as an educator to teach me more about G-theory than I ever thought I would know. Herb, many thanks for your detailed eye and thoughtful suggestions for this research. You provided a perspective that truly strengthened the quality of this work. Donna, I appreciate the high standards that you set for me (and all students). Additionally, I want to thank you for your generous support of the research study that made this dissertation possible. I hope that you found the outcome to be well worth CARS' investment.

I also want to acknowledge the Center for Faculty Innovation for helping to fund the faculty who participated in the APT ratings. Along with CARS' support, CFI's generosity provided a pathway for this research to move forward and I am grateful to them for their support of assessment.

To the faculty and fellow students who I have worked beside in CARS: thank you for sharing your talents with me. I continue to be amazed by the knowledge, humility, and talents that you all possess and hope that our relationships continue for many years.

I am grateful to have had the support of my family during this journey. To my parents, Richard and Sue Orem, I will be forever grateful for the love, encouragement, and guidance you always provide me. Your love and support made it easier to focus on the task at hand and has helped make this process as smooth as possible. To my other parents, David and Karen Barnes, thank you for the general concern for my well-being during graduate school. You both provided a physical presence that I sorely needed at times.

Finally, I want to acknowledge my wonderful, compassionate, and patient wife, Sarah. Throughout this experience, you not only kept me looking forward to the future, but helped me stay grounded in the present. Thank you for the sacrifices you've made during this process, and know that the support I received from faculty and fellow students in the program was outmatched by the support I received from you at home.

Table of Contents

List of Tables

List of Figures

Abstract

Meta-assessment, or the assessment of assessment, can provide meaningful information about the trustworthiness of an academic program's assessment results (Bresciani, Gardner, & Hickmott, 2009; Palomba & Banta, 1999; Suskie, 2009). Many institutions conduct meta-assessments for their academic programs (Fulcher, Swain, & Orem, 2012), but no research exists to validate the uses of these processes' results.

This study developed the validity argument for the uses of a meta-assessment instrument at one mid-sized university in the mid-Atlantic. The meta-assessment instrument is a fourteen-element rubric that aligns with a general outcomes assessment model. Trained raters apply the rubric to annual assessment reports that are submitted by all academic programs at the institution. Based on these ratings, feedback is provided to programs about the effectiveness of their assessment processes.

Prior research had used Generalizability theory to derive the dependability of the ratings provided by graduate students with advanced training in assessment and measurement techniques. This research focused on the dependability of the ratings provided to programs by faculty raters. In order to extend the generalizability of the meta-assessment ratings, a new fully-crossed G-study was conducted with eight faculty raters to compare the dependability of their ratings to those of the previous graduate student study. Results showed that the relative and absolute dependability of two-rater teams of faculty ($\rho^2 = .90$, $\Phi = .88$) were comparable to the dependability estimates of two-rater teams of graduate students. Faculty raters were more imprecise than graduate students in their ratings of individual elements, but not substantially.

Based on the results, the generalizability of the meta-assessment ratings was expanded to a larger universe of raters. Rater inconsistencies for elements highlighted potential weaknesses in rater trainings. Additional evidence should be gathered to support several assumptions of the validity argument. The current research provides a roadmap for stakeholders to conduct meta-assessments and outlines the importance of validating meta-assessment uses at the program, institutional, and national levels.

Chapter One

Introduction

The institutional effectiveness and accountability movement changed the landscape of higher education in many ways. In part, the need for institutional leaders to demonstrate the worth of higher education has led many U.S. academic degree (e.g., A.A., B.A., B.S., M.A, and Ph.D.) programs to engage in learning outcomes assessment. As part of this process, program faculty often produce an assessment report in which programs identify what learning objectives students are to achieve, the methodology employed to measure the objectives, the results, and the ways in which the program used the results to improve. Although changes to academic programs are often made based on anecdotal observations by faculty members or because of requests by upper-level administrators, an ever-growing number of academic programs use their assessment results to make data-driven decisions regarding curricular and instructional improvements. Ideally, however, the processes by which these decisions are made should be rooted in sound theory and incorporate appropriate methodology. Otherwise, faculty and administrators may come to incorrect conclusions about the degree to which their students are learning. Thus, to improve the quality of data that drive these decisions, universities are well-served to evaluate the processes by which their students are assessed. These processes—known as meta-assessments—are the focus of this research.

Meta-Assessment

In essence, meta-assessment is the process of evaluating assessment. Ory (1992) was the first scholar to use the term meta-assessment in a higher education context, articulating that the assessment field is inherently linked to that of evaluation.

Specifically, Ory believed that assessment scholars could draw upon the field of meta-evaluation—the evaluation of evaluation—to craft the procedures for meta-assessment. Similar to meta-assessment, the term meta-evaluation was first used to refer to a plan for evaluating educational products (Scriven, 1969 as cited in Stufflebeam, 2000). Current meta-evaluators (e.g., Stufflebeam, 2000; 2001) claim that it is critical for evaluators to engage in meta-evaluation to ensure that audiences make educated decisions—based on accurate information—about the quality of products or programs.

In order to evaluate higher education assessment processes, Ory (1992) suggested that assessment professionals use the Joint Committee on Standards for Education Evaluation's 30 *Standards for Evaluation of Educational Programs, Projects, and Materials* (Joint Committee, 1981). Grouped into four categories, the standards addressed the usefulness of the evaluation to its audience (Utility); the feasibility and cost effectiveness of the evaluation (Feasibility); the degree to which the evaluation was conducted legally and ethically (Propriety); and the accuracy of the information produced from the evaluation (Accuracy). Because Ory (1992) believed strongly that the field of evaluation was extremely relevant to understanding assessment, he saw these standards as a clear framework for conducting meta-assessment.

Ory (1992) is among several scholars to recognize the importance of conducting meta-assessment as part of institutional effectiveness (Bresciani, Gardner, & Hickmott, 2009; Hatfield, 2009; Palomba & Banta, 1999; Suskie, 2009; Walker, 1999). Bresciani, Gardner, and Hickmott (2009) added to the literature by distinguishing between meta-assessment at the program and the institutional levels. Typically, program-level meta-assessments enable practitioners to judge the processes by which individual academic programs produce and use assessment results. Usually, program-level meta-assessments

focus on judging the quality and appropriateness of certain core components of assessment such as the program objectives, methods, results, and the uses of data to support decisions (Bresciani, Gardner, & Hickmott, 2009; Fulcher, Swain, & Orem, 2012). In contrast, institution-level meta-assessments tend to evaluate the degree to which assessment is systemically conducted within an entire college or university. Institutional meta-assessments may focus on broader aspects of assessment such as the degree to which upper administration supports assessment by providing resources, the existence of learning outcomes at various institutional levels, or the use of results for budgetary decision-making. (Bresciani, Gardner, & Hickmott, 2009; Suskie, 2009).

There are relatively few examples in the literature of meta-assessment at either the institutional or program level. In some of the early references to meta-assessment, scholars were more interested in identifying the factors that contributed to strong assessment practice, instead of evaluating the assessment processes of a specific program or institution. Researchers with the California State Higher Education System, for example, conducted a meta-assessment to identify successful assessment characteristics of various colleges and universities (California State University—Long Beach Institute for Teaching and Learning, 1993). Their findings suggested that institutions practicing quality assessment often had solid faculty and administrator support. Additionally, strong assessment was conducted at institutions that employed personnel with advanced statistical and measurement abilities. These early meta-assessments provided foundational knowledge about effective assessment strategies; however one may consider them to be more exploratory in nature, and outside the purview of the present research.

Program-Level Meta-Assessments

Applied examples of program-level meta-assessments are particularly difficult to find in the literature. For instance, although Ory (1992) made a strong case for using the evaluation standards as a foundation for conducting meta-assessment, there is no published research that incorporates his model. This lack of research could be due to the fact that, although Ory's philosophy regarding the use of evaluation standards to inform meta-assessment processes makes sense in theory, implementing this process at the program level poses incredible logistical challenges. Much like meta-evaluations tend to focus on one specific evaluation, Ory likely envisioned meta-assessments evaluating one assessment process. Given the large number of academic programs at many institutions, Ory's model is too complex to implement given the limited resources often devoted to assessment practice.

In the only published example of an applied meta-assessment, assessment practitioners at Marquette University used a rubric to identify improvements in program assessment across multiple years (Fong Bloom, 2010). Fong Bloom's work illustrates how institutions can use a meta-assessment rubric to evaluate and track assessment processes over time. And, although Fong Bloom provides few details about how these scores were validated, her research highlights the potential uses of a program-level meta-assessment.

Despite few examples of applied meta-assessment in the literature, many schools do in fact conduct program-level meta-assessments. In fact, a recent study found that over 50 institutions use rubrics or checklists to evaluate the veracity of program assessment processes (Fulcher, Swain, & Orem, 2012). The results indicate that while little research on these measures exists, assessment practitioners across the United States are actively

engaging in meta-assessment and institutions are recognizing the value of this work. Nevertheless, more research is needed to deem whether these rubrics and larger meta-assessment processes yield valid results.

The current research adds to the small existing meta-assessment literature base through an examination of the reliability of academic program-level meta-assessment scores across multiple populations of raters. In addition to examining the consistency of rater scores across programs, the variability of ratings on individual elements can also be derived. This research will serve as a model for other institutions wishing to estimate sources of systematic variability in their own meta-assessment processes, thereby providing direction for an understudied, yet increasingly prevalent, aspect of assessment.

Generalizability Theory

As a necessary precursor to any validity argument involving a performance assessment such as a rubric it is important to determine whether reliable ratings can be attained. If, after reading the same information, two raters come to different conclusions regarding a program's assessment process, then any subsequent validity argument is greatly weakened. In other words, when inconsistencies exist among raters, it becomes unclear what is actually being assessed by the instrument, and how the subsequent scores can be used. Thus, it is important to demonstrate that scores on a meta-assessment rubric are consistent across raters. Several techniques are available to estimate reliability; however, Generalizability theory (G-theory) is the most relevant strategy for this study for four reasons: 1) it produces an estimate of inter-rater reliability to help answer pertinent research questions regarding the consistency of meta-assessment ratings; 2) in addition to estimating the systematic error attributable to raters (i.e., inter-rater reliability), one can also estimate additional sources of error (e.g., difficulty of elements)

and their interactions that may impact scores; 3) alternate conditions can be tested using G-theory (e.g., D-studies) to provide practical and well-informed alternate designs intended to improve reliability; and 4) it can produce a reliability estimate around a particular point on a scale (i.e., for absolute decisions). A richer discussion of G-theory follows in chapter two.

Literature Review Framework

Although dozens of institutions use meta-assessments to evaluate program-level assessment processes, applied examples and empirical studies of specific measures are dramatically absent from the literature. This paper begins to address the lack of literature by examining, in depth, the ratings of one such meta-assessment instrument. In order to provide the reader with appropriate context and rationale for this research, chapter two begins with a brief historical review of the factors leading to the assessment and accountability movement. Following this discussion, various definitions of assessment will be shared and the characteristics of good assessment will be identified. At this point in the literature review, the focus will shift from assessment to meta-assessment, in particular, how it is defined, its role in the larger assessment process, its connection to the field of meta-evaluation, and its current use at both the institution and program-level in higher education.

After the general history of assessment and concept of meta-assessment have been discussed, the scope of the paper narrows further, and the specific meta-assessment process in question is introduced and described within a common meta-evaluation framework. The discussion then turns to the meta-assessment rubric used in this specific process, and the validity argument is presented to support the use of this instrument. Specifically, the validity argument integrates Kane (1992) and Benson's (1998)

frameworks to present the current evidence supporting the uses of the meta-assessment rubric. Subsequently, the areas of the validity argument needing further exploration are identified, including the areas of research covered specifically within this paper. Because the current evidence supporting the uses of the meta-assessment rubric requires an understanding of G-theory, sufficient attention will be devoted to explaining this measurement technique within the broader discussion of validity.

Chapter Two

Background

The Assessment Movement

The current assessment movement began largely as a result of the *Involvement in Learning* report (Ewell, 2002; Study Group on the Conditions of Excellence in American Higher Education, 1984). The report was a response to calls by the National Institute for Education and the United States Department of Education to determine ways of improving undergraduate higher education in America. It offered 27 suggestions based on three primary recommendations: 1) Student involvement in learning needed to increase, 2) Clear expectations and high standards about what could be accomplished had to be shared by both students and institutions, and 3) Evaluation and assessment had to be a central part of academic learning. The intent of the report, according to its authors, was to "contribute to the national discussion and action on improving quality in postsecondary education" (p. vii). Assessment was seen, in large part, as the vehicle to drive higher education's quality enhancement.

Although *Involvement in Learning* was written with the goal of pedagogical and curricular improvement in mind, the recommendations mostly aligned with governmental mandates for accountability within institutions of higher education (Ewell, 2002). Thus, the beginning of the movement was also a response to these mandates for evidence that higher education was not only making good use of federal dollars, but was still capable of driving the U.S. economy (Erwin, 1991; Ewell, 2002; Rossman & El-Khawas, 1987; Shavelson, 2010). Many early scholars viewed assessment as being the institution's response to these two philosophically different, yet inter-connected factors: the desire to improve, and responding to external calls for accountability (Erwin, 1991; Ewell, 1988;

Jacobi, Astin, and Ayala, 1987). There is little doubt that these two factors are still a driving force of contemporary assessment.

Defining assessment. As the movement progressed, one of the challenges to the initial practitioners was defining assessment. Early scholars produced several similar, yet varied definitions, many of which were derived from the more mature concept of evaluation. As a tool for decision-making, evaluation has been defined as a formal collection of information that is used as a basis for making judgments (Stufflebeam, 1968). The second edition of the Joint Committee on Standards for Educational Evaluation (1994) defined evaluation simply as "the systematic investigation of the worth or merit of an object" (p. 3), whereas the third edition (2011) expanded the definition to include:

> The systematic investigation of the quality of programs, projects, subprograms, subprojects, and/or any of their components of elements, together or singly for purposes of decision making, judgments, conclusions, findings, new knowledge, organizational development, and capacity building in response to the needs of identified stakeholders leading to improvement and/or accountability in the users' programs and systems ultimately contributing to organizational or social value. (p. xxv)

In short, evaluation is the process of making judgments or decisions about an object or process, based on some level of systematic evidence. The basis for many of the current definitions of assessment stem from this concept.

Fields outside of education have used assessments to gauge the effectiveness of their programs (Erwin, 1991). However, the following discussion focuses on the use of

assessment in postsecondary educational settings. The educational literature identifies several general definitions of assessment. First, assessment has been viewed as any process of gathering evidence about the impact of higher education (Boyer & Ewell, 1988; Davis, 1989; Ory, 1992). Under this definition, all aspects of a university can be assessed, from the services provided by campus safety, to content knowledge gained from a senior seminar. This definition is closely aligned with the concept of evaluation, in which the "object" is education. Scholars in this school of thought may also see assessment as being most useful when it is done as part of a large-scale program to test student performance at the institution level (Ewell, 2002). In order to assess student learning broadly, schools can use an abundance of standardized instruments to benchmark student learning at their institutions relative to others. From such instruments, summary statistics can be produced quickly and efficiently (Ewell, 2002; Shavelson, 2010).

Although most assessment scholars would likely agree that assessment is, in part, the process of investigating the impact of higher education, this definition leaves much to be desired. For one, it does not differentiate between outcomes and outputs, an important distinction when evaluating higher education's effectiveness. Outputs (e.g., retention rates, graduation rates, or fundraising dollars) help institutions evaluate many important functions, but they do not provide any information about the cognitive, developmental, or affective impact colleges have on student learning. For instance, a high graduation rate is important to a university's reputation, but this statistic provides no information about what a student has learned in his or her time at the institution. In response to this shortcoming, scholars have advocated for a second definition of assessment, in which the focus of higher education effectiveness rests squarely on measuring student learning

(Davis, 1989; Erwin, 1991; Jacobi, Astin, and Ayala, 1987; Marchese, 1987; Rossman & El-Khawas, 1987). This definition makes a clear distinction between evaluation, which does not require the inclusion of student learning, and outcomes assessment, which is, at its core, the collection of evidence to support claims surrounding what college graduates know, think, or do as a result of their experience.

Outcomes assessment, the process of determining the skills, knowledge, and abilities that students gain as a result of college, has been widely adopted by many assessment practitioners as an optimal approach to gathering evidence about the impact of college on student learning. However, this process still has its roots in evaluation. Tyler (1950) viewed evaluation as a goal-oriented process, in which behavioral goals are specified, data are collected using instruments chosen to measure said goals, and then the data are analyzed, interpreted, and used for improvements.

Many prominent assessment scholars have embraced this model, using it to further delineate and define the components of the assessment process (Erwin, 1991; Palomba & Banta, 1999; Suskie, 2009). This general model describes a process that closely resembles Tyler's goal-oriented approach to evaluation. Beginning with the creation of measurable goals and outcomes, professionals develop appropriate instruments that adequately measure the knowledge, skills, and behavior expected of students. Assessment practitioners then implement processes to systematically collect information, and identify methods of analysis to interpret the data. The results are used to make changes to programs, courses, departments, or any number of areas that might improve future iterations of data.

As Erwin (1991) argued, Tyler's approach to evaluation has been adopted by many institutions as a model for assessment largely because his goal-oriented approach

aligned with the National Governor's Association (Alexander, Clinton, and Kean, 1986) philosophy that institutions should have well-defined missions and outcomes. Although Erwin made this argument close to twenty years ago, more recent literature suggests that the National Governor's Association still maintains its belief in the importance of an outcomes-driven philosophy regarding higher education (National Governor's Association [NGA], 2007; Reindl & Reyna, 2011).

By defining assessment broadly, it is apparent why institutions view it as an important process for facilitating student growth—it is a systematic process by which strengths and weaknesses of programs, courses, or institutions can be identified and improved to positively affect student learning. To effectively evaluate assessment, however, one must go beyond a general conceptual definition and explore the specific qualities that characterize a strong assessment process.

Characteristics of Outcomes Assessment

Intended as a general resource for assessment practitioners, Suskie (2006) compiled a summary of good assessment practices from a variety of assessment scholars and professional organizations. Suskie's summary provided the basis for this discussion about good assessment, but other sources were consulted in order to compile the most comprehensive review possible. When discussing the components of a strong assessment process, it is important to first determine the level at which assessment is to take place. Certain practices apply more to institutional processes, whereas other characteristics more appropriately inform sound programmatic assessments.

At the institution level, strong assessment occurs when various stakeholders view the process as important and assessment efforts are supported across all levels of the institution (e.g. faculty, administrators, students; American Association of Higher

Education [AAHE] 1993; Banta, 2002; Council of Regional Accrediting Commissions, 2004; Huba & Freed, 2000). Additionally, assessment endeavors should be cost-effective given the resources available (Bresciani, 2003; Driscoll & Cordero De Noriega, 2006; Huba & Freed, 2000; Suskie, 2009). Furthermore, institutions must ensure that their assessment processes are sustainable. The sustainability of assessment processes can be strengthened by establishing assessment offices, engaging faculty, and setting realistic expectations (Bresciani, 2003; Greater Expectations Project on Accreditation and Assessment, 2004; Palomba & Banta, 1999; Suskie, 2009).

These characteristics of a strong institutional-level assessment process are certainly important, but they are also difficult to assess. Peterson and Einarson (2001) conducted a study of institution-level assessment, but the results were comparative across institutions and not designed to address specific universities' assessment practices. Additionally, accrediting agencies have developed instruments to identify strengths and weaknesses of assessment processes beyond academic programs (e.g., Middle States Commission on Higher Education, Western Association of Schools and Colleges). It is evident that while identifying aspects of strong assessment practices at the institutional level is possible, it is more challenging to examine objectively the extent to which an institution possesses these characteristics.

Narrowing the scope, multiple books and articles explore the characteristics of good assessment for academic programs. The vast majority of these materials express the importance of identifying clear learning objectives, and many of these authors suggest that the objectives should be tied to the institutional mission (Driscoll & Cordero De Noriega, 2006; Erwin, 1991; Palomba & Banta, 1999; Suskie, 2009). Furthermore, quality assessment establishes a link between the experiences and activities programs

offer, and programs' goals and objectives (American Association of Colleges and Universities [AACU], 2002; Australian Universities Teaching Committee, 2002; Suskie, 2000). To assess objectives, many authors suggest using multiple, direct measures in order to minimize any negative effects resulting from limitations with a single instrument (Australian Universities Teaching Committee, 2002; Greater Expectations Project on Accreditation and Assessment, 2004; Suskie, 2000).

Good assessment also involves appropriately and thoroughly examining, sharing, and using results. Though also true at the institutional level, Erwin (1991) states that good program assessment focuses first on improvement and second on accountability. Palomba and Banta (1999), among others, offer similar advice, although they argue that at the very least programs need to have a clear purpose for their assessment, using both formative and summative assessments to measure their outcomes (AACU, 2002; AAHE, 1993; Australian Universities Teaching Committee, 2002; Banta, 2002; Greater Expectations Project on Accreditation and Assessment, 2004). Perhaps most importantly, assessment results need to be trustworthy. That is, there should be some assurance that the assessment process leads to relatively accurate, fair, and useful information (Australian Universities Teaching Committee, 2002; Eder, 1999 as cited in Suskie, 2006; National Council on Measurement in Education, NCME, 1995; Suskie, 2009). These characteristics are all aligned with the general assessment model described earlier.

There are also important qualities of strong assessment that do not align with any specific model. For one, program (and institutional) assessment is not episodic, but rather, it is an ongoing process that builds on past iterations of results (AACU, 2002; AAHE, 1993; Banta, 2002; Greater Expectations Project on Accreditation and

Assessment, 2004; Steen, 1999). Additionally, assessment measures should target various cognitive and developmental stages of student learning and development (Australian Universities Teaching Committee, 2002; Erwin, 1991; Greater Expectation on Accreditation and Assessment, 2004; Palomba & Banta, 1999), not focus solely on lower-order thinking skills or basic levels of student development.

In summary, good program assessment aligns with stages in the assessment model proposed by scholars like Erwin (1991) and Palomba and Banta (1999). To practice good assessment, programs should have clearly stated and well-reasoned learning outcomes that are tied in some way to the larger goals and mission of the institution. The purpose of the assessment should be clear, both to those designing it and to the students assessed. If the purpose is summative, those constituents who will be directly impacted need to understand the decisions made as a result of the assessment. Multiple instruments should be used to assess the objectives. These measures are intentionally chosen or designed to assess cognitive, developmental, and attitudinal growth in students, and should be diverse in nature in order to supplement any individual measure's limitations. Courses, experiences, and activities should be linked directly back to objectives to strengthen the connection between what is expected of students and what is provided to them in order to meet those expectations. The methods used to assess students should be fair, systematic, and intentional, and the results should be interpreted accurately in order to strengthen the inferences made from the results. Additionally, the assessment process should be sustainable, and intentions should be made to continue to refine and revisit the entire cycle so that the program's effectiveness is strengthened with each iteration.

Meta-Assessment

These components of good assessment are all important in determining higher education's impact on student learning. However, one characteristic remains to be discussed: the proper, systematic evaluation of the assessment process itself, or as it will be referred to in this paper, meta-assessment (AACU, 2008; Banta, 2002; Eder, 1999 as cited in Suskie, 2006; Hatfield, 2009; Huba & Freed, 2000; Ory, 1992; Palomba & Banta, 1999).

Peters (2005) defined meta-assessment as the "deliberate examination of the elements, basic conditions (necessary and sufficient), and needs of a thing (service, event, system and so on) that transcend particular instances of that thing" (p. 347). McDonald (2010) carried this definition further, arguing that, with meta-assessment, one "is questioning the tenets of the program instead of simply accepting what one is provided with" (p. 120). Simply, meta-assessment is the process of evaluating assessment practice, and it can provide crucial information about the veracity of a program or institution's assessment processes. Without periodic review of assessment processes, flawed methods—among other things—may lead programs or institutions to use results inappropriately for important decisions. Flawed assessment can do more harm than strong assessment does good; therefore, it is advantageous for any institution to have systems in place to monitor and evaluate the quality of institution and program level assessment processes.

The current research focuses on the effectiveness of meta-assessment strategies regarding academic program assessment. Conducting assessment that yields trustworthy and useful results is challenging, requiring diligence and careful design. Meta-assessment can help academic programs discern the strengths and weaknesses of their

assessment. These programs can then improve upon their existing processes to produce a higher quality of assessment. As we begin to examine various meta-assessment processes to uncover best practices, we should first revisit the connection between evaluation and assessment.

Meta-Evaluation. The fields of evaluation and assessment are distinct; however, Ory (1992) articulated that the world of assessment is inextricably linked to that of evaluation. As early assessment practitioners attempted to respond to external calls for results, they overlooked the ways in which evaluation theory could have aided them in the process (Ory, 1992). Specifically, the field of meta-evaluation, or the evaluation of evaluation, was an area that Ory believed assessment scholars could draw upon for guidance. In 1969, Michael Scriven used the term "meta-evaluation" to refer to a plan for evaluating educational products (Stufflebeam, 2000). Stufflebeam (2000, 2001) has since defined meta-evaluation as:

> the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation's utility, feasibility, propriety, and accuracy and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide the evaluation and publicly report its strengths and weaknesses (p. 96; p.183).

Meta-evaluation, Stufflebeam argued (2000), is imperative to ensuring that evaluators deliver effective results that allow audiences to make good decisions about the quality of products and programs.

Stufflebeam (2001) identified the ten main steps of meta-evaluation, located in Table 1. As these steps illustrate, meta-evaluations align generally with the stages of the assessment cycle described earlier in the chapter: questions to be evaluated are framed, a

process to collect and analyze data is specified, and the results are interpreted to relevant stakeholders. Additionally, several of the steps indicate logistical details that should be included in the process (e.g. framing a contract). More importantly, the process includes identifying the standards on which the evaluation system in question will be judged. Stufflebeam (2001) argued that the evaluation should be judged against the standards of the specific evaluation field. For example, the evaluation of a particular educational program might use the Joint Committee's Program Evaluation Standards (1981; 1994; 2011).

Table 1

The Ten Main Steps of Meta-evaluation

1. Determine and arrange to interact with the meta-evaluation's stakeholders
2. Establish a qualified meta-evaluation team
3. Define the meta-evaluation questions
4. Agree on standards to judge the evaluation system or particular evaluation
5. Frame the meta-evaluation contract
6. Collect and review pertinent available information
7. Collect new information as needed, including, for example, on-site interviews, observations, and surveys
8. Analyze the qualitative and quantitative information and judge the evaluation's adherence to the selected evaluation standards
9. Prepare and submit the needed reports
10. Help the client and other stakeholders interpret and apply the findings

Practical examples of meta-evaluation are common in the literature (Fitzpatrick, 2004; Scott-Little, Hamann, & Jurs, 2002), Many of these examples emphasize the importance of aligning evaluations with the appropriate standards. For example, Cooksy (1999) conducted a meta-evaluation of a curriculum delivery for middle schools using the American Evaluation Association's guiding principles (AEA, 1994 as cited in AEA, 2004) and the Joint Committee's evaluation standards (1994). Grasso (1999) compared

the same Joint Committee's standards to an evaluation of a reading program for war veterans, determining whether various standards and guiding principles were met. Many meta-evaluators rely on professional standards to identify strengths and weaknesses of specific evaluation processes, thereby supporting or invalidating the inferences evaluators hope to make about certain programs or products. Additionally, examples of meta-evaluation can be found in non-education fields, such as healthcare and business, further illuminating the fact that the field of evaluation is much broader than that of assessment (Cottarelli & Escolano, 2004; Smits & Champagne, 2008).

Connecting meta-evaluation to meta-assessment. Early assessment practitioners of the 1990s faced many of the same challenges that evaluators experienced in the 1960s; specifically, they questioned how to perform quality assessment, and more importantly, how to *know when* they were performing quality assessment (Ory, 1992). Ory attempted to help assessment practitioners answer these questions through the same techniques of meta-evaluators; namely, by aligning assessment processes and procedures with certain standards. Because no agreed upon assessment standards existed in the early 1990s (nor do they exist today), Ory connected the two fields by explaining the practical applications of the Joint Committee on Standards for Educational Evaluation's *Standards for Evaluation of Education Programs, Projects, and Materials* (1981; 1994) for assessment. Using these 30 standards as the basis for his argument, Ory (1992) believed that assessment processes could be evaluated along the same guidelines. These 30 standards are grouped into four categories: 1) Utility Standards (i.e., the assessment is useful), 2) Feasibility Standards (i.e., the assessment is cost effective and reasonable), 3) Propriety Standards (i.e., the assessment is legal and ethical), and 4) Accuracy Standards (i.e., the assessment produces accurate information).

Although Ory's (1992) philosophy regarding the link between meta-evaluation and meta-assessment makes theoretical sense, the practical implementation of his meta-assessment at the program-level falls short. Imagine, for instance, an institution with over 100 academic programs requiring assessment reports that addressed all 30 standards. Not only would it be challenging for program faculty to write such extensive reports, but the resources needed for assessment professionals to review the reports and provide effective feedback would be enormous. Given this scenario, one can quickly see how Ory's concept of meta-assessment at the academic program level would be practically untenable.

Since Ory (1992), additional scholars have emphasized the importance of meta-assessment in conducting sound assessment (Bresciani, Gardner, & Hickmott, 2009; Hatfield, 2009; Palomba & Banta, 1999; Suskie, 2009; Walker, 1999). Much like assessment, processes exist both for institutions and for academic programs; Bresciani, Gardner, and Hickmott (2009) made the distinction between institutional- and program-level meta-assessment. Program-level meta-assessments allow practitioners "to evaluate the processes by which individual programs produce assessment results" (Orem & Fulcher, in review). Typically, these meta-assessments address the program objectives, methods, results, and the uses of data to support decisions (Bresciani, Gardner, & Hickmott, 2009; Fulcher, Swain, & Orem, 2012). Institution-level meta-assessments tend to examine instead the degree to which "assessment is systemic" (Orem & Fulcher, in review). Institutions may be assessed on the level to which leaders support assessment work, results drive budgetary decisions, or learning outcomes exist at the institution and department level (Bresciani, Gardner, & Hickmott, 2009; Suskie, 2009).

Examples of meta-assessment. Only a few scholarly references describe applications of meta-assessment at any level of higher education. One of the earliest examples of a broad-based meta-assessment in higher education was completed by Johnson, Prus, Anderson, and El-Khawas (1991). The researchers investigated the extent to which campuses had conducted assessment activities and had assessment offices. Similarly, Peterson and Einarson (2001) reviewed a multitude of institutional assessment practices to determine how different types of universities (e.g. research, comprehensive, liberal arts) conducted different kinds of assessment and used results in varying ways.

Additionally, the Fund for the Improvement of Postsecondary Education sponsored a meta-assessment of public institutions in California, identifying common factors of good assessment. The results of the study revealed that faculty and administrative support for assessment was imperative. Further, the presence of assessment practitioners with strong statistical and measurement backgrounds also enhanced the quality of the institution's assessment (California State University—Long Beach Institute for Teaching and Learning, 1993). These types of exploratory meta-assessments evaluated assessment trends across multiple institutions, however, rather than investigating the quality of assessment within specific institutions. This research will focus largely on meta-assessments that are institutionally-specific, particularly at the academic program level.

Examples of program-level meta-assessments in the literature are rare. In one of the only published examples of an applied program-level meta-assessment, Fong Bloom (2010) discussed the process implemented by her university to evaluate assessment. Using a rubric to rate programs on five aspects of assessment (learning outcomes; assessment measures; assessment results; faculty analysis and conclusions; and actions to

improve learning and assessment) Fong Bloom demonstrated how a meta-assessment could help school administrators and program faculty identify improvements to program assessment across several years. Because of the meta-assessment, institutional leaders were able to identify programs that were meeting or exceeding expectations regarding assessment; aggregate results across several years to target specific aspects of the assessment process requiring systemic improvement; and more effectively collaborate with the faculty of those programs not meeting the university's assessment standards in order to improve their processes. Since implementing the meta-assessment process, Fong Bloom (2010) concluded that more programs were successfully completing the entire assessment cycle.

Aside from Fong Bloom (2010), little research exists that focuses on applied program-level meta-assessment; however, the *practice* of program-level meta-assessment is quite prevalent in higher education. As noted earlier, Fulcher, Swain, and Orem (2012) found over 50 institutions that use rubrics or checklists to evaluate aspects of program assessment. Using one of these meta-assessment rubrics, raters might provide specific scores for areas of assessment such as the quality (or presence) of objectives, curriculum mapping, assessment instruments, data collection processes, and use of results. These components illustrate the aspects of assessment certain schools deem most important to evaluate, or most practical given the resources available to them. Given the benefits of conducting program-level meta-assessment, it is certainly encouraging that many institutions are using meta-assessment instruments to evaluate their assessment processes. However, although the research by Fulcher, Swain, and Orem (2012) indicates that dozens of institutions currently use meta-assessment instruments, little is known about

the quality of these processes. That is, the practice of meta-assessment is outpacing its scholarship.

This research will add to the current literature by investigating crucial empirical and theoretical questions regarding program-level meta-assessment. Specifically, what does an effective meta-assessment process look like, and more importantly, what type of evidence must exist to label a meta-assessment process effective in the first place? To answer these questions, meta-assessment will be conceptualized and discussed using the process adopted at James Madison University (JMU) to evaluate program-level assessment processes. Because a specific framework for conducting meta-assessment is still absent, JMU's meta-assessment process will be explained within the broader 10-step framework of meta-evaluation (Stufflebeam, 2000). There are two purposes for using a meta-evaluation framework to introduce JMU's process. One, the framework provides a clear organizational structure by which to fully introduce and describe JMU's meta-assessment process. Two, by using a meta-evaluation framework to introduce a meta-assessment process, the similarities between the two fields become clearer.

After explaining JMU's meta-assessment process, the validity argument for the uses of scores from the meta-assessment instrument will be presented. Exploring the important assumptions and questions related to the validity of JMU's meta-assessment process—and to the reliability of scores ascertained from the instrument used to evaluate program assessment—will be a crucial part of this paper's contribution to the assessment literature.

Meta-assessment at James Madison University: The Assessment Progress Template

The Assessment Progress Template (APT) process is an emerging application of program-level meta-assessment at James Madison University. The APT is an assessment

report completed annually by academic programs (See Appendix A). A rubric is used to evaluate the reports (i.e., APT rubric), enabling reviewers to provide diagnostic feedback to each academic program about the quality of its assessment practice (Fulcher, Sundre, & Russell, 2009; see Appendix B). What follows are the ten steps of a meta-evaluation (Stufflebeam, 2000; see Table 1 for the full list of steps) along with a description of how the evaluation of the APTs mirrors this basic framework.

Step 1: Determine and arrange to interact with the meta-evaluation's stakeholders. During this initial step of the meta-evaluation process, the target audience of the meta-evaluation is established and contact with them is made (Stufflebeam, 2000). Eventually reports will be written about the evaluation process, and these reports should be written to the appropriate populations. During the APT meta-assessment, initial meetings are held between assessment experts and program heads, assessment coordinators, and other institutional stakeholders at various times throughout the year. These meetings have been formal gatherings with assessment coordinators in which the meta-assessment process is discussed and the expectations for APT reports are shared with faculty. The APT meta-assessment process is intended to be formative, and thus, the eventual meta-assessment reports are written to address program faculty and assessment coordinators who are in charge of their programs' assessment processes.

Step 2: Establish a qualified meta-evaluation team. As the stakeholders in the meta-evaluation are being identified, qualified evaluators must be chosen to provide accurate and appropriate feedback to programs about their evaluation processes. During the APT meta-assessment process, graduate students with specific training in assessment practice are solicited as part of their assistantship requirements to review the APTs. In addition to their assessment experiences, these raters receive training on the standards

used to judge the reports. The raters also provide qualitative feedback to programs to complement the ratings. This feedback includes recommendations for improving weaker areas and praise for strong components of the program's assessment process. The entire rating process is overseen by a faculty member with over seven years of experience conducting and researching meta-assessments.

Step 3: Define the meta-evaluation questions. Stufflebeam (2000) states that certain fundamental questions must be asked when conducting a meta-evaluation: (1) to what extent does the evaluation meet the audience's need for evaluative information and (2) how well does it meet the requirements of a sound evaluation (p. 102). In other words, does the evaluation have worth, and is it meritorious (Stufflebeam, 2000)? Clearly, these are fundamental questions that speak to the validity of the evaluation, specifically the uses of the evaluation scores. The program should be using the evaluation results to make substantive changes to either its program, or perhaps its evaluation process. If the evaluation results are not used except to fulfill some requirement, then the evaluation process does not have much worth. Similarly, if the methodology of the evaluation is grossly flawed, or the uses of the results are not appropriate given the findings, then the evaluation lacks merit, and the impending validity argument is weak. In either case, the meta-evaluation should speak to these issues, addressing both the worth and the merit of the evaluation in question (Stufflebeam, 2000).

The APT meta-assessment process is designed to answer these same questions concerning individual programs' assessments. Program faculty members are provided with feedback to help them use the results effectively and they are given advice to help them determine the extent to which their results can be trusted. Thus, the APT meta-

assessment process is designed to help programs evaluate the worth and merit of their assessment processes. Clearly, the degree to which the meta-assessment process helps programs meet these objectives requires more attention than what is given here. However, as the larger validity argument for the uses of the APT meta-assessment is presented, greater attention will be paid to exploring these important issues.

Step 4: Agree on standards to judge the evaluation system or particular evaluation. During this stage of meta-evaluation, various standards are determined. The evaluation is then judged according to these standards. Here, the *evaluation* is the program's assessment process, which is detailed in the APT report submitted yearly. Trained raters then use the APT rubric to judge the particular assessment process provided in the APT report. Each element of the rubric constitutes an agreed upon *standard* by which sections of the APT report are evaluated. The rubric consists of 14 elements that are rated on a four point scale: 1 – Beginning, 2 – Developing, 3 – Good, 4- Exemplary (see Appendix B). These elements correspond to six areas of the assessment process and include criteria like the presence of clear and precise objectives; sound data collection and research design; and evidence of programmatic and assessment improvements based on results. Using this four point scale, programs are rated on the degree to which various elements of their assessment process align with the behavioral anchors. In addition to assigning the element a numerical score, raters also provide feedback to the programs about each of the elements, identifying strengths and weaknesses of their assessment processes. These standards were developed by assessment practitioners and approved by an assessment advisory council that agreed upon the elements and the anchors.

Step 5: Frame the meta-evaluation contract. This step of the meta-evaluation process incorporates each of the previous two stages (Stufflebeam, 2000). The meta-evaluation contract between the meta-evaluators and the program should outline the questions to be answered and the standards by which the process will be judged (p. 99). Additionally, the contract provides details about needed information, the report deadlines, and authorship of the reports. A similar process is followed in the APT meta-assessment process, although no official contract is signed. Programs are given a deadline by which to submit their APT, and in return, the Center for Assessment and Research Studies (CARS), the office coordinating the APT assessment process, names a general date by which the feedback will be provided. This timeline is approved by both an Assessment Advisory Council as well as an Academic Council of Deans. Furthermore, department heads and CARS representatives discuss the dissemination of the APT feedback reports, determining who within the university receives access to the feedback.

Step 6: Collect and review pertinent information. After the details of the meta-evaluation contract are determined, the meta-evaluation team collects information pertinent to the process. Stufflebeam (2000) refers to this step as the "desk-review" (p. 99). In the case of the APT meta-assessment, programs submit the APT assessment report (see Appendix A) via an electronic submission system shortly after the end of the spring semester. Ideally, this report addresses all six areas of the APT report that is then evaluated using the 14-element rubric (see Appendix B). In other words, the assessment report will discuss objectives, a curriculum map, methodology, results, communication with stakeholders, use of results, and the more granular elements within these areas. Programs may also include attachments of assessment results or any additional documentation to inform the raters about their assessment process.

Step 7: Collect new information as needed, including, for example, on-site interviews, observations, and surveys. The meta-evaluation team may determine, after reviewing the initial information, that more is needed to fully inform their overall meta-evaluation argument. In this stage, the team goes to the site of the evaluation to conduct interviews and may even observe the evaluation process in order to achieve a fuller understanding of the evaluation in question. For the APT meta-assessment, this step is not entirely transferable. After the program submits its report, the raters must use only what has been provided to assign the program scores. This limitation is largely due to resource constraints and the sheer size of the APT meta-assessment process. Remember that most meta-evaluations are conducted on a single program, not 120. Because raters must read and evaluate multiple reports in a short amount of time, they are unable to contact programs to "fill the gaps." Ideally, programs would be able to provide clarification on areas of their assessment process that raters find unclear.

Even more so, raters would be able to observe the program's assessment process in person, including faculty meetings, test administrations, and discussions about the uses of results. In this sense, the scores on the APT rubric would be a more accurate reflection of programs' assessment processes because they would be based on direct observations. However, raters must use the information provided in the report to judge the quality of the program's assessment. Therefore, it is imperative that programs submit an accurate and representative report of their processes. Multiple resources are provided to faculty to aid them in writing a complete, well-structured report such that these information gaps are limited. Further, programs receive feedback, not only about the quality of their assessment, but also about the clarity and organization of their reports. Because APTs are collected and evaluated each year, programs can help fill gaps by clarifying and

improving their reports with each reporting cycle, leading to more accurate representations of their assessment processes over time.

Step 8: Analyze the qualitative and quantitative information and judge the evaluations' adherence to the selected evaluation standards. After all pertinent information has been collected, the meta-evaluation team analyzes the findings and determines the extent to which the program's evaluation follows the agreed-upon standards. During the APT meta-assessment, the raters (i.e. the meta-evaluation team) receive a full day of training on the rubric. During this training, raters receive an overview of all the elements and a chance to practice scoring example APTs in order to calibrate their scores with each other. Raters are then paired together and assigned a random sample of APTs to judge. Each member of a rater team initially assigns scores independently but both members periodically review their scores together (post initial rating) in order to discuss discrepancies and calibrate their ratings. The numerical scores are analyzed using descriptive statistics, inferential ANOVA procedures, and advanced generalizability theory techniques. These procedures help the meta-assessors identify overall strengths and weaknesses of program assessment, evaluate the amount of program growth year over year, and analyze the generalizability of scores across raters and items.

Step 9: Prepare and submit the needed reports. After meta-evaluators analyze the data, they prepare the agreed upon reports for the client. During the APT meta-assessment process, programs receive a feedback report from CARS that contains the 14 element scores, an average item score across all elements, and the comments from the raters. These reports go to the assessment coordinators and department heads. Deans receive a summary report containing the numerical average, across all 14 elements, of each program in their colleges. The provost receives all of the dean-level reports. .

Step 10: Help stakeholders interpret and apply the findings. After the reports have been sent to the client, the final stage of the meta-evaluation process entails making sure the client understands the findings. Often, programs receive low scores and copious feedback on the APT rubric because they do not understand what is being asked of them. To address these misunderstandings, programs receive a variety of resources to help them interpret the ratings and identify ways to improve scores on the next iteration of the process. Program Assessment Support Services (PASS) is an office that provides assessment support to academic programs. This office is staffed by graduate students with skills in assessment consultation. These consultants provide help in the form of individual consultations, workshops on various aspects of the assessment process, data analysis, and report writing. Full-time faculty members with assessment and measurement expertise also provide advice to programs about their assessment processes and facilitate the interactions between program faculty members and the PASS office. These resources are available year round for faculty to help encourage constant reflection and action regarding assessment practices.

As this section of the chapter has illustrated, the meta-assessment process at JMU is designed to provide academic programs with a thorough, accurate evaluation of their assessment processes. One of the most visible and political components of this process is the APT rubric, specifically the scores that are shared with various stakeholders and the process by which those scores are determined. Therefore, it is of utmost importance that the APT rubric scores be trusted and are meaningful to faculty and administrators.

The Concept of Validity

To address the trustworthiness of the APT scores, it is important to understand how they are used by stakeholders. Scores on the APT rubric are used in four ways: (a)

they provide practitioners with a means to evaluate the quality of academic program assessment processes, (b) they help program faculty identify strengths and weaknesses of their assessment procedures, (c) they provide a common metric for aggregated and disaggregated descriptive statistics, and (d) they can be used to gauge the effectiveness of an assessment office (Fulcher, personal communication, 2010). However, what evidence exists to support these uses? Do the scores a program receives on the rubric accurately reflect its assessment practices? Are the current uses of the APT rubric warranted? What are the implications of making decisions based on inaccurate data? In order to trust the interpretations made from the rubric scores—and instill faith in stakeholders that using rubric scores will lead to informed data-driven decision making—these questions must be addressed through a validation process.

Our focus now shifts to an examination of the validity evidence that currently exists to support the uses of the APT rubric. Beginning with a discussion of validity, the various frameworks used to form a validity argument will be described. Then, the current validity argument for the uses of the APT rubric scores will be given. This section will conclude with a discussion of the current weaknesses in the validity argument and future research that must be conducted to strengthen areas where the validity argument is limited.

Validity is the degree to which evidence supports the intended uses and interpretations of test scores (Messick, 1989). Originally, different forms of validity were thought to fulfill distinct purposes (Messick, 1989; Shepard, 1993). For example, content validity was used to determine the degree to which a test's content adequately covered the domain of interest (e.g., are the questions on a math test representative of and relevant to the defined domain of math ability?). Criterion validity referred to the degree to which

test scores were related to other similar measures of the same domain (e.g., do scores on a depression inventory relate to scores on a measure of anxiety in a way we would expect?). Construct validity was needed when determining the extent to which the test interpretations supported or aligned with the construct(s) in question (Shepard, 1993).

As the concept of validity evolved, construct validity came to be seen as the unifying factor incorporating all other types of validity (Messick, 1989). Messick argued that content validity was technically not a form of validity because the test's content was not inherently used to interpret scores. Furthermore, criterion validity, Messick reasoned, was too specific to make a case for validity on its own. Thus, Messick claimed that any evidence used to support a test's interpretations was fundamentally construct validity. Since Messick's article, other scholars have supported this unified concept of construct validity (Shepard, 1993).

In conceptualizing the idea of construct validity, Cronbach and Meehl (1955) introduced the term "nomological network" to refer to a construct's place within a larger conceptual landscape of all related constructs and hypotheses. Thus, all relevant theories that have been formulated to explain the construct and its relationship with other variables exist within this nomological network. Validity evidence is gathered by testing the hypotheses about these relationships in an effort to support or refute the conclusions made about the construct (Benson, 1998). Thus, the various forms of validity (e.g., content, criterion) were seen as supplements to one another, instead of alternatives, that informed the larger construct validity argument (Messick, 1989). Additionally, because construct validity is conceptualized as *the degree* to which evidence exists to support test interpretations, scholars have argued that one can never gather enough validity evidence

to fully support an intended use (Benson, 1998; Cronbach, 1989; Messick, 1989; Shepard, 1993).

If the validation process is never-ending, how does one begin to tackle such a seemingly insurmountable task? Cronbach (1989) outlined two "programs" of validation. The *weak program* required researchers to provide evidence of correlations between the construct of interest, measured by the test in question, and some related construct. Thus, these "raked together" correlations were all that was needed to satisfy basic requirements for the validity of test score interpretations (Cronbach, 1989, p. 155). The *strong program* was more rigorous, and involved the researcher explicitly identifying the construct, describing the theoretical hypotheses used to understand the construct, and gathering evidence to support or reject the stated hypotheses (Cronbach, 1989). Thus, this program of construct validation involved gathering both theoretical and empirical evidence to support test score interpretations and testing rival hypotheses to falsify alternative theories (Benson, 1998).

Kane's (1992) Interpretive-Argument Approach

The strong program of validation has been conceptualized in a variety of ways by different researchers. Kane (1992) created the argument-based approach to validation, which he described as being similar to Cronbach's strong program (p. 534). Kane (1992) claimed that test score interpretations always incorporate an interpretive argument, in which conclusions are made about test scores in the form of statements and decisions (p. 527). According to Kane (1992), there are three criteria that must be met when making an interpretive argument. First, the argument must be clearly stated. What conclusions are to be drawn from the scores? What assumptions are made about the procedures and the scores that support these conclusions? How is the construct defined? Second, the

interpretive argument should be coherent. In other words, do the specified assumptions lead to reasonable conclusions? Is the argument practical, and does it flow logically? Can the assumptions be tested theoretically or empirically? Third, the assumptions must be plausible. In most cases, assumptions cannot be proven true, but evidence can be gathered to refute rival hypotheses and to support the hypotheses made about the assumptions. In order to meet this third criterion, various forms of validity evidence must be used to test these hypotheses (Kane, 1992). Thus, much like Cronbach's original description of the strong program, Kane's approach involves a clear statement of the argument one wishes to make about the scores, a description of the assumptions (i.e., hypotheses) required to make the argument, and evidence that the assumptions have been met (or not shown to be false).

This method has many strengths; however, the process by which a particular set of scores is validated is not very obvious within Kane's structure. A researcher may know that assumptions need to be developed and tested, but may be unsure of which assumptions to test first so that the validity argument has a logical flow. To understand how a validity argument might follow such a pattern, we turn to Benson's three-stage approach to construct validation (1998).

Benson's (1998) Three-Stage Approach

Benson (1998) used Nunnally's (1978) framework to operationalize the strong program of validation (Cronbach, 1989). This framework consists of three stages to construct validation: (a) the substantive stage, (b) the structural stage, and (c) the external stage.

The substantive stage. The substantive stage, according to Benson (1998) involves defining the construct of interest, both theoretically and empirically. The

theoretical domain consists of all the ways that the construct (e.g., assessment) has been defined previously in the literature. The theoretical domain is then operationalized in the empirical domain, which consists of all the possible observable variables and types of instruments (e.g., writing samples, observed behavior, selected response test) that could reasonably measure the defined construct (Benson, 1998). Thus, the purpose of the substantive stage is to gather evidence to support both the definition of the construct and the ways in which it is intended to be measured (Benson, 1998). During this stage, it is important to ask whether or not the definition of the construct aligns with the intended uses of the instrument.

The structural stage. After the construct has been theoretically defined and a measure has been developed that operationalizes the theoretical definition, the researcher moves into the structural stage, where validity evidence must be gathered that supports the internal consistency of the scores. During this stage, researchers examine the items on the instrument used to measure the construct to determine the extent to which the items represent the construct (Benson, 1998). Benson outlined several popular methods for gathering validity evidence in this stage, notably confirmatory factor analysis and generalizability theory. During this stage, the researcher should ask whether or not the items on the instrument produce results that can be used to interpret the focal construct. This stage is most often associated with the dimensionality and reliability of test scores and ratings.

The external stage. After sufficient evidence has been gathered to support the substantive and structural stage of validation, the researcher must demonstrate that the construct in question covaries with other constructs in previously theorized ways (Benson, 1998). The scores on the instrument in question (e.g., the APT rubric) should

not only demonstrate a positive relationship with measures of similar constructs, but they should also produce results that diverge from unrelated constructs. This evidence can be gathered through processes such as group differentiation and correlations with similar and distinct measures (Benson, 1998). Here, one needs to ask whether the scores on the instrument can be used to support the construct's theorized relationship to other variables in the nomological network. Because Benson's framework occurs in stages, without evidence in the substantive and structural stages to support the uses of the test, it is inappropriate to conduct analyses in the external stage.

The Validation Argument for the Uses of the APT Rubric Scores

By studying both Kane (1992) and Benson's (1998) approaches to framing a validity argument, one begins to see how these two models can be integrated to establish evidence supporting the proposed uses of the APT rubric. Following Benson's framework provides a logical flow to the validity argument, starting with the substantive stage and progressing through the structural and external stages. Throughout this process, Kane's argument-based approach can help formulate the assumptions (i.e., hypotheses) that, by addressing, will provide evidence to support the intended uses of the meta-assessment. To illustrate a validity argument that incorporates both approaches, consider the assumptions and evidence that currently support the uses of the APT rubric scores (see Table 2).

Table 2

List of Assumptions in the APT Rubric Validity Argument

Assumptions
Substantive Stage
1. The rubric adequately covers the breadth and scope of the assessment construct
2. The rubric is an appropriate method to measure the assessment construct
3. The assessment report writer accurately represents the quality of the program's assessment process in the APT
Structural Stage
4. Raters produce consistent scores relative to one another
5. Raters consistently agree on a program's score relative to the behavioral anchors on the APT rubric
External Stage
6. Higher scores on the APT rubric reflect better practices in assessment.
7. The rubric scores are related to other measures of assessment or related constructs in the hypothesized direction

Current evidence for the substantive stage. The foundation to any validity argument begins with identifying the focal construct. Messick (1989) argued that the term "test score" is used generically when referring to validity. In reality, a validity argument can be made about any observed consistency, such as performance tasks or other assessment methods. The APT rubric is one such example of a non-traditional form of assessment in which scores are interpreted to reflect the quality of academic programs' assessment processes. Thus, the focal construct to be defined and operationalized is *learning outcomes assessment*. The content of the APT rubric and the resulting scores are used to represent this construct. Therefore, to provide evidence for the substantive stage, the theoretical and empirical domains of the APT rubric must be identified and

justified. Thus, three assumptions must be supported to begin building the case for validation.

Assumption one. The rubric adequately covers the breadth and scope of the assessment construct. Evidence must be gathered that shows how the rubric's development was grounded in theory relevant to assessment practices, and that steps were taken to minimize construct irrelevant variance and underrepresentation.

In support of assumption one, the APT rubric was created by assessment practitioners with expertise in the realm of assessment and measurement procedures. The fourteen rubric elements were developed to correspond with a general outcomes assessment model championed by other experts in the field (Erwin, 1991; Palomba & Banta, 1999). The rubric developers consulted pertinent literature that supported the general structure of the rubric; however, they also relied on personal experiences from their work as assessment practitioners to craft the various descriptions of the elements (Fulcher, personal communication, 2010). The rubric was also reviewed and approved by additional professionals with assessment expertise. Thus, the theoretical domain of the construct was developed from expert knowledge and prior literature. Furthermore, the current fourteen element rubric evolved from a six element rubric. Elements were added to several areas of the rubric, particularly the methodology section. The rubric was expanded to its current length in an effort to minimize construct underrepresentation and address areas of construct irrelevance.

Assumption two. The rubric is an appropriate method to measure the assessment construct. Thus, evidence should show that other forms of assessment were considered and that the current rubric is logical given these other measures.

Prior to the APT rubric, assessment reports were evaluated with a checklist that focused largely on the six main components of the assessment cycle. However, this evaluation format did not provide much information to faculty about the strengths and weaknesses of their assessment, leading many stakeholders to criticize the quality of feedback they were receiving (Fulcher, personal communication, 2010). In response to these faculty criticisms and at the recommendation of an assessment advisory council (AAC), the APT rubric was created to improve on the previous methodology.

Since the development of the APT rubric, researchers have conducted reviews of other forms of program-level meta-assessments in higher education. Noted earlier, Fulcher, Swain, and Orem (2012) found over 50 institutions that use some form of rubric to evaluate assessment processes; the rubric was by far the most common assessment format used by the institutions conducting meta-assessments. This research provides evidence, albeit retroactively, that supports the use of a rubric for program meta-assessment purposes. The prevalence of meta-assessments that use rubrics combined with the evolution of the APT rubric demonstrates sufficient support for this second validity assumption.

Assumption three. The assessment report writer accurately represents the quality of the program's assessment process in the APT. Much like the decision to test a student's critical thinking via a performance assessment, multiple choice test, or self-report survey, the assumption here is that an assessment report is an appropriate way for a program to demonstrate its assessment practice. In essence, violation of this assumption could occur if a program is conducting stellar assessment work, but the person tasked with writing the APT turns in an unorganized, incomplete report. This scenario may lead raters to score the program low on the rubric, not because the program practices poor

assessment, but because the report itself does not provide accurate information. One can think of this issue as a method effect. Thus, the assumption in question is that scores on the rubric are reflective of the program's assessment, and are not confounded by irrelevant constructs such as a poorly written report.

To demonstrate support for this assumption, evidence must show that the content of the submitted APT report is an accurate reflection of the program's actual practices. To meet this assumption in an ideal world with limitless resources, someone with assessment expertise would observe the processes by which all academic programs conduct their assessment. These observations would then be compared to the content of the submitted report. Clearly, this evaluation method is unrealistic. Therefore, the APT serves as the program's best representation of its assessment process. Using this logic, three issues may occur in which error would be added into the scores. One, the report may be poorly written, and thus, low ratings reflect the rater's inability to decipher the author's writing. The scores are then confounded with the error introduced by poorly communicated processes. Two, the report's author may omit important information that would have resulted otherwise in a more accurate, and likely higher score. In this case, the rubric scores might be thought of as a lower bound to the program's true representation of the assessment construct. Third, instead of omitting information, the author might provide inaccurate information, resulting in a higher (or possibly lower) score. These are important points to consider and future work should focus on satisfying this assumption. However, the evidence supporting assumptions one and two provide strong support for the substantive validity of the APT rubric scores, allowing for the validation argument to shift into the second stage of Benson's framework.

Current evidence for the structural stage. As the validity argument transitions into the structural stage of Benson's model, evidence must be gathered to support two assumptions. In order to determine the degree to which these two assumptions are met, a brief discussion of generalizability theory (G-theory)—the method used to examine these assumptions—is required. Additionally, because the current study focuses heavily on exploring these assumptions, it is necessary to provide proportionally more detail here about previous work that provides the foundation for this research.

G-theory measures reliability using a framework called dependability. Shavelson and Webb (1991) define dependability as "the accuracy of generalizing from a person's observed score on a test … to the average score that a person would have received under all possible conditions that the test user would be equally willing to accept (the Universe of Admissible Observations; UAO)" (p. 1). Within the context of the APT scores, the UAO includes all possible academic programs that could be rated by the rubric, all feasible elements that could be included on the rubric, and all possible raters that could be trained to rate the APTs.

One purpose of G-theory is to determine the extent to which the score a program receives from the two raters is consistent with the average score it would earn if it were rated by all raters that could feasibly be trained on the rubric (e.g., faculty members, graduate students, assessment professionals) on all potential elements that comprise the assessment construct. This average score is known as the universe score and is analogous to the true score in classical test theory. As outlined by Shavelson and Webb (1991), a person's observed score on an assessment—or in this case a program's observed score on the APT rubric—is comprised of several sources of variability. The numerical values that represent these sources of variability are referred to as variance components.

One source of variability, known as the object of measurement, comprises the systematic differences in real ability between programs' assessment processes (σ^2_p). This variability is synonymous with true score variability in classical test theory. The second source of variability exists within the facets of the measurement design. Facets are the sources of systematic error within a particular design. In the APT measurement design, variability among raters, due to the inherent subjectivity in human ratings (σ^2_r) is a source of systematic error and is considered a facet. Thus, after scoring all rubric elements for all programs, some raters may be systematically harsher or more lenient than other raters. Another potential facet in the APT measurement design is the differences in difficulty across all programs and raters among the fourteen elements (σ^2_i). That is, after every program has been scored by every rater, some elements on the rubric may be generally harder for all programs to do well, whereas other elements may generally be much easier.

Another source of variability involves the interactions between both the facets and the object of measurement. Regarding APT ratings, certain raters may be harsher than others when scoring specific elements, in which case an interaction between rater and element exists (σ^2_{ri}). Some elements may be more difficult for some programs to do well relative to others (i.e., a Program x Element interaction; σ^2_{pi}). Raters may also score certain programs more harshly than others, wherein program APT scores will be rank ordered differently depending on the rater (i.e., a Program x Rater interaction; σ^2_{pr}). The final source of variability discussed by Shavelson and Webb (1991) is the residual variance ($\sigma^2_{pri,e}$). It consists of the random error captured in the observed scores (e.g., overlooked information within a report), other sources of systematic error not modeled in the design, or a combination of both.

Facets can be either random or fixed. Random facets refer to sources of error—rater opinions or differences in item difficulty—that are sampled from a larger universe of equally qualified raters or equally difficult items. Fixed facets refer to an aspect of a study in which all possible elements from the population are included in the measurement. For example, fixed facets may be used in cases where every rater that could reasonably be expected to score an assessment is being used, where one does not wish to generalize beyond the items or raters being used, or where the rubric elements that are scored effectively represent the entire construct of interest.

The sources of error can be used to calculate the total amount of relative and absolute variance within a particular measurement design. When considering the variability in APT scores, relative variance illustrates the precision with which programs are rank ordered similarly across raters and elements. Absolute variance indicates the precision with which program ratings, across raters and elements, are consistent with the rubric's behavioral anchors. The sources of error that contribute to the relative variance in APT scores are the Program x Rater interaction (σ^2_{pr}), Program x Element interaction (σ^2_{pi}), and the Program x Element x Rater plus Random Error interaction ($\sigma^2_{pri,e}$). Sources of error that factor into the absolute variance of APT scores include the $\sigma^2_{pr,}$ $\sigma^2_{pi,}$ and $\sigma^2_{pri,e}$ variance components as well as the Rater effect (σ^2_{r}), Element effect (σ^2_{i}), and the Rater x Element interaction (σ^2_{ri}). A list of these variance components and their notations is found in Table 3.

Table 3

Variance Components in the Fully Crossed Design of APT Ratings

Variance Component Notation	Description of Variance Component
σ^2_p	Program (object of measurement)
σ^2_r	Rater facet
σ^2_i	Element facet
σ^2_{pr}	Program x Rater interaction
σ^2_{pi}	Program x Element interaction
σ^2_{ri}	Rater x Element interaction
$\sigma^2_{pri,e}$	Program x Rater x Element interaction plus Random Error

Two types of analyses are conducted within a G-theory analysis. The first, called a G-study, calculates the variance components of a single observation for the object of measurement and the facets. That is, how far would a single, randomly chosen rater's average score across all programs and elements be from another, randomly chosen rater's average? Or, if we were to randomly select one element from the universe of potential elements, how close would its average score be from another element's average—across all raters and elements—that was also selected randomly from the element universe?

The variance component information from the G-study is used to complete the second analysis, known as a decision study (D-study). D-studies are used to estimate relative and absolute variance. Given a specific design (two raters scoring all fourteen APT elements), one is able to estimate the accuracy of a program's observed score to its universe score. Additional D-studies can be conducted in which the conditions of measurement are altered (e.g., using three raters instead of two, or scoring 20 elements instead of 14), thus affecting the relative and absolute dependability estimates. These

alternate D-studies allow researchers to determine the conditions needed to achieve a desired level of dependability; they can be quite useful when allocating the appropriate amount of resources to score a performance assessment.

To calculate both the relative and the absolute variances, the relevant variance components from the G-study are each divided by the number of levels of the facets in the D-study and summed. Equations (1) and (2) illustrate how to calculate the relative (equation 1) and absolute (equation 2) variances of design involving a rater and element facet.

$$\sigma^2_{\text{Rel}} = \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{pri,e}}{n'_r n'_i} \quad [1]$$

$$\sigma^2_{\text{Abs}} = \frac{\sigma^2_{r}}{n'_r} + \frac{\sigma^2_{i}}{n'_i} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{ri}}{n'_r n'_i} + \frac{\sigma^2_{pri,e}}{n'_r n'_i} \quad [2]$$

The variances obtained from Equations 1 and 2 can be used to derive relative and absolute dependability estimates for particular measurement designs. These estimates are represented by the G- and phi-coefficients. G-coefficients (ρ^2) measure the proportion of variance in a sample that can be attributed to universe score variance compared to relative observed score variance (see Equation 3). Essentially, the G-coefficient provides information about the accuracy with which programs' assessments are ranked. Phi-coefficients (Φ) estimate the proportion of universe score variance compared to the absolute observed score variance (see Equation 4). That is, what is our confidence that a program really scored a "2"? Both estimates range between zero and one, with larger values indicating higher levels of dependability in the design. Because phi-coefficients take into account additional sources of error (e.g., Raters x Elements) they are lower than, or equal to, G-coefficients, which are calculated using only the relative error variance.

$$\rho^2 = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{REL})} \quad [3]$$

$$\Phi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{ABS})} \quad [4]$$

G-studies can be either fully-crossed or nested. A fully-crossed design—one in which all raters rate all elements for every program—provides the most stable estimates of variance components, and enables one to consider a large number of possible design structures for future APT ratings (Brennan, 1992). In most cases, a fully-crossed design is preferred to a nested design (e.g., raters do not score every element; raters do not rate all programs). Although many G-studies use nested designs, certain variance components will be confounded with other variance estimates, making it extremely difficult, if not impossible, to separate out the variance attributable to specific sources of error (Brennan, 1992; VanLeeuwen, 1997).

As with any performance assessment that uses human raters, it is important to demonstrate that programs are rated consistently. If, after reading the same report, multiple raters arrive at drastically different conclusions about the quality of a program's assessment process, the validity argument regarding the interpretation of those scores is significantly weakened. Therefore, scores on any meta-assessment rubric must be consistent across raters. G-theory is a powerful technique for examining meta-assessment ratings because, unlike other reliability estimation methods that approximate error from only one source (e.g., Cronbach's Alpha — items only; or Cohen's Kappa – raters only), G-theory enables the researcher to identify the consistency of scores across raters, while simultaneously estimating additional sources of systematic variance (e.g., difficulty and consistency of items). Further, G-theory's capability to calculate absolute

dependability estimates allows the researcher to determine the consistency of meta-assessment scores relative to the rubric's criteria, which, in the case of the APT meta-assessment, is more useful than looking at relative dependability alone.

When considering potential validity assumptions that fit in the structural stage of Benson's (1998) framework, the relative and absolute consistency of rater scores are arguably the most important. The evidence gathered in the structural stage of Benson's (1998) framework should establish support for the relative and absolute consistency of these ratings across the elements on the APT rubric. Thus, G-theory is a useful tool for determining the level of consistency among raters who score the APT rubric. In the structural stage, there are two assumptions that require supporting evidence, both dealing with consistency of the ratings. Each of these assumptions will be defined separately, but because evidence for them was gathered at one time, the support for both assumptions will be discussed simultaneously.

Assumption four. Raters produce consistent scores relative to one another. That is, a rater may be harsher or more lenient than another rater, but this harshness is consistent across all programs and elements. This means that although the overall rubric scores assigned to programs may differ by rater, the overall rank order of program scores remains the same across all raters. Thus, evidence must exist that the relative variance of the program scores is minimal.

Assumption five. Raters consistently agree on a program's score relative to the behavioral anchors on the APT rubric In order to meet this assumption, not only would raters rank order programs the same way, but they would also converge around the same value on the rubric scale (e.g., 3.4). Therefore, the absolute variance among the facets should be small.

A small variance component associated with the Rater x Program variance component ($\hat{\sigma}^2_{pr}$) provides evidence that raters are consistent in how they rank order programs, thereby supporting assumption four. Support for the fourth assumption is also bolstered by a large G-coefficient, which estimates the relative dependability of ratings. Assumption five can be supported with evidence that raters not only rate programs consistently between each other, but that scores align with the standards outlined on the rubric itself. To strengthen this fifth assumption, the phi-coefficient can be estimated for a number of different designs (e.g., two raters scoring fourteen elements, three raters scoring fourteen elements), providing information about the ideal design needed to reach adequate dependability. Because the phi-coefficient takes into account all sources of systematic variance within a certain design, it provides a way for researchers to estimate how close a certain observed score is to its universe score.

There is ample evidence to support the fourth and fifth assumptions, all of which was collected simultaneously. In the summer of 2010, a fully-crossed G-study was conducted on the rubric scores from the 2009-2010 APTs (Orem & Fulcher, in review). The results of the G- and D-studies from the 2010 study indicated strong support for both the relative and absolute consistency of APT ratings (See Table 4). Specifically, assumption four, which addresses relative consistency among raters, is best supported by a high G-coefficient, which would indicate that raters were capable of rank ordering programs in a similar fashion. The results of the 2010 study using a two-rater, 14-element universe of generalization and considering the Element facet fixed yielded a G-coefficient of .92 (Orem & Fulcher, in review), which is more than an adequate estimate of relative dependability. Additionally, the Program x Rater variance component ($\hat{\sigma}^2_{pr}$

= .0228), which is a major indicator of the relative consistency among raters, accounted for only 2.4 percent of the total variance in the G-study. This finding suggests that very little of the overall variability in ratings was due to raters rank ordering programs differently.

As evidence for assumption five, which focuses on absolute consistency among raters, the phi-coefficient for the two-rater, 14-element universe of generalization—in which the Element facet was considered fixed[1]—was high ($\hat{\Phi}$ = .91; Orem & Fulcher, in review). This level of dependability indicated that, given the 14-element rubric, raters' scores were consistent relative to the behavioral anchors on the rubric. That is, two graduate student raters, when chosen randomly from the universe of possible raters, were able to consistently come to agreement on the scores they gave to programs using the APT rubric.

As further support for assumption five, the G-study revealed a small variance component for the rater facet ($\hat{\sigma}^2_r$ = .0055; Orem & Fulcher, in review). This number indicates that, on average, one rater's score across all elements and all programs is only .005 points different than the grand mean. Given the scope of the rating scale (one to four points), this difference in overall ratings is quite small and strongly suggests that the graduate student raters who scored the rubrics could provide ratings consonant with the behavioral anchors on the rubric, as opposed to some self-determined scoring criteria.

[1] The rationale for fixing the Element facet will be provided in the Chapter Three

Table 4

2009-2010 APT Ratings Using the Fully Crossed Design: Contribution of each Facet to Score Variance

Source of Variation	$\hat{\sigma}^2$	G Study[a] $nr = 1$ $ni = 1$	% Total Variance	D Study[b] $nr = 2$ $ni = 14$	% Total Variance
Program (p)	$\hat{\sigma}^2_p$	0.2171	23.6	0.2310	91.3
Rater (r)	$\hat{\sigma}^2_r$	0.0055	.6	0.0032	1.2
Element (i)	$\hat{\sigma}^2_i$	0.2540	27.8	---	---
pr	$\hat{\sigma}^2_{pr}$	0.0228	2.4	0.0188	7.4
pi	$\hat{\sigma}^2_{pi}$	0.1961	21.3	---	---
ri	$\hat{\sigma}^2_{ri}$	0.0125	1.3	---	---
pri,e	$\hat{\sigma}^2_{pri,e}$	0.2088	22.8	---	---
$\hat{\sigma}^2_\delta$				0.0189	
$\hat{\sigma}^2_\Delta$				0.0221	
$\hat{\rho}^2$				0.92	
$\hat{\Phi}$				0.91	

Note. [a]Variance components if the element facet is random.
[b]Element facet is treated as fixed.

Although the G-study provided a strong case for the relative and absolute dependability of the APT scores, it is necessary to mention the generalizability of the raters. At most institutions with similar models of assessment evaluation, faculty members provide the ratings and the feedback. The raters of the APT rubric, however, are all graduate students with assessment experience. Thus, they possess a unique lens through which to provide feedback. Because of their assessment knowledge, their perspectives likely increase the amount of rater consistency in the scores. However, most institutions do not use raters with this specific expertise. Therefore, one may ask whether the results of the APT rating method can be generalized to a population of academic

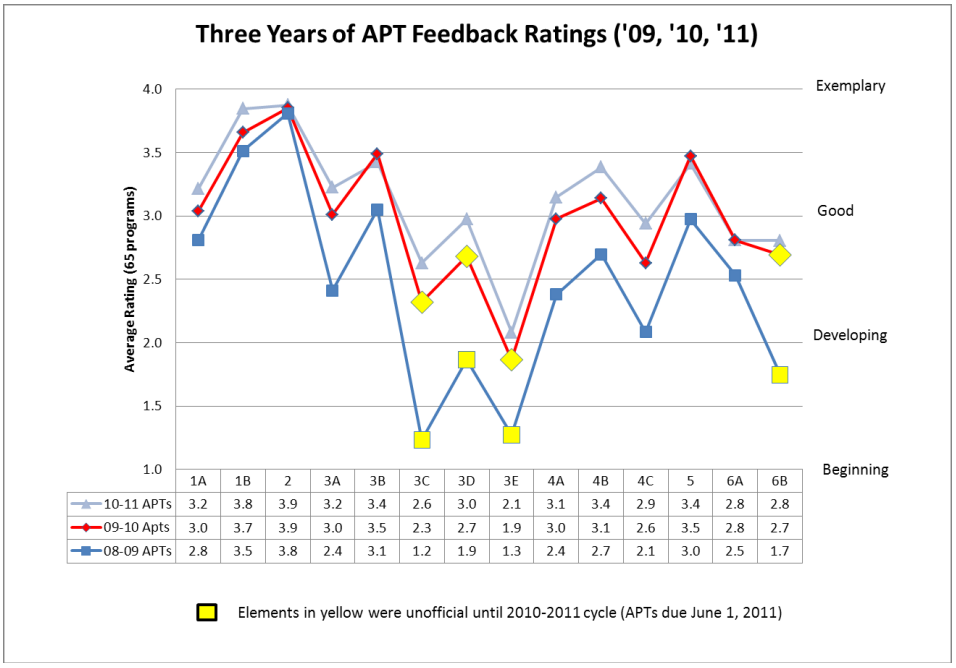
programs beyond the scope of James Madison University. Furthermore, one could argue that faculty may be more appropriate raters from a political perspective. Essentially, peer review of assessment would be consistent with faculty governance.

Current evidence for the external stage. Remember that validation is a never-ending process. Thus, although limitations certainly still exist to fully support either the substantive or structural stages of program validation, it is evident that to some degree, the APT rubric has been developed with both theoretical and operational considerations, and raters consistently interpret the rubric. Given this support of the substantive and structural stages thus far, attention can now turn to the external stage of Benson's (1998) validation argument. In this stage, two additional assumptions must be met.

Assumption six. Higher scores on the APT rubric reflect better practices in assessment. Thus, a program with a good assessment process should score substantially higher on the rubric than a program with a mediocre or limited assessment procedure. This type of validity is called known groups validity. This assumption also holds true when examining the assessment processes within the same program over multiple years. In other words, the scores on the rubric should reflect real changes in a program's assessment and correct for possible omissions regarding the assessment process from one year to the next.

A comparison of APT scores of programs from the 2008-2009, 2009-2010, and 2010-2011 academic years provides the strongest evidence in support of assumption six. In theory, as programs understand the assessment cycle and are given additional resources by which to conduct assessment, their scores should improve. Figure 1 illustrates the mean scores of the 14 elements across programs for three academic years (2009 to 2011). It is evident from Figure 1 that programs, year over year, did improve in

every element—in some areas, drastically. Thus, one could take these results to mean that programs learned from the APT feedback and improved in the areas in which they were weakest.



	1A	1B	2	3A	3B	3C	3D	3E	4A	4B	4C	5	6A	6B
10-11 APTs	3.2	3.8	3.9	3.2	3.4	2.6	3.0	2.1	3.1	3.4	2.9	3.4	2.8	2.8
09-10 Apts	3.0	3.7	3.9	3.0	3.5	2.3	2.7	1.9	3.0	3.1	2.6	3.5	2.8	2.7
08-09 APTs	2.8	3.5	3.8	2.4	3.1	1.2	1.9	1.3	2.4	2.7	2.1	3.0	2.5	1.7

Figure 1. Trends of element mean scores from the APT rubric. Source: The Center for Assessment and Research Studies, Harrisonburg, VA

Additional evidence, however, needs to demonstrate that these score differences are true reflections of improvements to program assessment. The information provided in Figure 1 provides initial support that the rubric can identify improvement in assessment processes. The average element scores of programs with ratings across all three years were charted, with growth occurring year over year across almost all elements. One may question, however, whether this growth is due to real changes in the assessment process, or is due to increased leniency among raters. To ensure that this growth is due to substantive improvements, programs with large changes in assessment scores year to year are flagged by experienced raters, and the reports are compared directly to the previous year's reports to determine whether the changes in scores are real,

or due to rater interpretation. This quality control process effectively eliminates a vast number of potential rating errors that could artificially inflate or deflate the overall average element scores. Thus, the increase in scores shown in Figure 1 is the result of actual improvements to programs' assessment processes, as demonstrated by the APT reports. This data provides validity evidence in support of assumption six.

To further explore the reasons scores on the APT rubrics increased year to year, eleven assessment coordinators from programs with the most drastic improvements were interviewed during spring 2011 to determine exactly what they did differently to warrant such increases. The results indicated that the vast majority of these programs (10 of 11) made substantive changes to their processes based largely on feedback they received from assessment practitioners (Fulcher et al., 2011). If these findings generalize to all programs that demonstrated some improvements in scores, the argument could be made that the increase in meta-assessment scores year over year is due in large part to real changes in programs' assessment processes.

Assumption seven. The scores on the rubric are appropriately correlated with other measures of the assessment construct, or other related constructs. To meet this assumption, it might be necessary to determine whether the scores on the rubric are positively correlated with other likely indicators of strong assessment performance (e.g. assessment coordinator's years of experience; the degree to which a program uses available resources; the motivation of the program's faculty to conduct assessment).

Data are limited to support assumption seven. Alternative indicators of strong or weak assessment have yet to be officially identified, and the data for these indicators must be tabulated and analyzed prior to conducting any statistical procedures. Additionally, certain indicators (e.g. the number of times a program consults resources)

might not be strong measures of assessment, calling to question the validity of these additional performance indicators. For example, a program may consult with an assessment liaison once a week, but never take any of the suggestions offered for improvement, and thus, their assessment may not improve, even with additional help. Instruments to measure the motivation of the assessment coordinators and their faculty to conduct assessment may provide one of the most plausible pieces of evidence to correlate with assessment scores, but these measures are yet to take shape. Future directions for this assumption will be covered in the discussion.

Strengthening the Validity Argument

Cronbach's (1989) strong program of construct validation is based on program evaluation theory (Shepard, 1993). With program evaluation theory, the areas of a program most critically in need of improvement are addressed first. Similarly, strengthening a validity argument requires paying attention to the assumptions most in need of additional evidence. To that end, Cronbach (1989) recommended determining the most pertinent validity questions to answer first, and then creating a process to address them. Several assumptions are still in need of additional validity evidence; however, greater evidence must be gathered that improves the generalizability of the APT ratings. Therefore, for the purposes of the current research, a closer examination of assumptions four and five warrants top priority.

Additional evidence for assumptions four and five. Evidence already exists to demonstrate the consistency of APT ratings when the raters are graduate students with advanced training in assessment. However, one may question whether the ratings can generalize beyond graduate students. That is, can other raters, specifically faculty, interpret the rubric consistently, with respect to other faculty as well as to the graduate

students with assessment training? Thus, what is the size of the universe to which the APT meta-assessment ratings can generalize? To answer this question and provide additional evidence in support of the structural stage assumptions four and five, faculty members – as opposed to graduate students - should be used to rate some, if not all of the APTs. This scenario is much more realistic for many institutions, and thus, the scores on the rubric must be able to reflect program assessment using raters similar to those at other universities. Graduate students specializing in assessment and measurement may provide expertise in the rating process above and beyond what faculty members could provide given similar resources. Although introducing expert opinion into the rating process is not necessarily bad, it could be argued that scores on the rubric should be consistent and consonant with the behavioral anchors regardless of the population (graduate students or faculty) used to provide them.

Purpose

The APT rubric is an emerging example of meta-assessment within higher education. As with any assessment instrument, the case must be built that the inferences made from the rubric's scores are valid. Given the need for additional validity evidence, this research serves two purposes. First, it can provide a framework by which other users of meta-assessments can validate their own processes. Second, the rubric itself has the potential to be used by institutions outside of JMU, improving the overall field of higher education assessment. To accomplish this second purpose, however, the ratings must be shown to be dependable when using a population of raters (i.e., faculty) more appropriate for a variety of institutions outside of JMU.

Research Questions

Therefore, this research provides additional validity in the structural stage of Benson's (1998) framework, where the generalizability of the ratings can be examined and improved. In this stage, scores have been dependable when using two graduate student raters scoring all elements. At most institutions, faculty members—and not graduate students—are the ones available to evaluate assessment reports. Thus, to improve the generalizability of the ratings, evidence that faculty members can dependably use the rubric must be provided.

This research uses a sample of faculty raters—with varying levels of familiarity with assessment—to answer several questions related to the validity of the APT rubric scores. First, are estimates of the relative and absolute dependability of faculty ratings adequate? That is, can faculty produce consistent scores relative to programs as well as to the rubric standards? Second, how similar are the dependability estimates of faculty raters compared to the estimates gathered from graduate student raters? Should these two groups produce similar dependability estimates, the case can be made that faculty and student raters are essentially interchangeable, important information to know when evaluating the interpretability of the rubric. Third, on average, are faculty raters harsher or more lenient than graduate student raters? Finally, which individual elements on the APT rubric have the largest standard errors? That is, on which elements do raters tend to disagree most? Do the elements that tend to produce rater inconsistencies among faculty mirror the elements that produced large standard errors among the graduate student raters? By answering these questions, this research will not only advance the effectiveness of the APT rubric for use at JMU, but it will further the literature in higher education regarding effective methods for program assessment evaluation

Chapter Three

Method

Measure

Every academic degree and certificate program at the university is required to submit an assessment report (the APT) annually. A 14-element, behaviorally-anchored meta-assessment rubric (the APT rubric) was used to evaluate the six major components of the APT (see Fulcher & Orem, 2010). The components of the APT—and subsequent meta-assessment rubric—were aligned with various stages of a cyclical, outcomes-based assessment model (see Table 5) : (a) student-centered learning objectives; (b) mapping learning experiences to the objectives; (c) methodology for collecting information; (d) results; (e) dissemination of results; and (f) the use of information for program and assessment improvements. These six components are further delineated into 14 elements. Raters score each element of the rubric on a four point scale (1—Beginning, 2—Developing, 3—Good, 4—Exemplary; half points are allowed). Ninety-seven APTs from the 2009-2010 academic year and 119 APTs from the 2010-2011 academic year were used in this study, representing 95% and 99% respectively of all university academic degree programs required to submit assessment reports over the two years.

Procedures

The rating process from 2009-2010 is described here in full for context, but also so that the reader recognizes the attempt by the researcher to provide a similar process during the 2010-2011 study. The 2009-2010 study is not an official part of the dissertation, but rather, it was part of a separate study that provided the foundation for this current body of research. Because this dissertation is greatly intertwined with the methodology of the previous study—the results from the two studies will in fact be

compared to answer multiple research questions—it is pertinent to state the 2009-2010 procedures here.

Table 5

The Stages and Elements of the Assessment Model Used for the APT Rubric

Stage	Element	Description
1. Specify Student Centered Learning Objectives	1a.	Clarity and specificity
	1b.	Orientation
2. Map Course/Learning Experiences to Objectives	2.	Course/Learning experiences are mapped to objectives
3. Systematically Evaluate Progress on Objectives	3a.	Relationship between measures and objectives
	3b.	Types of Measures
	3c.	Specification of desired results for objectives
	3d.	Data collection and research design integrity
	3e.	Additional validity evidence
4. Analyze Results of Program Assessment	4a.	Presentation of results
	4b.	History of results
	4c.	Interpretation of results
5. Documents How Results are Shared with Faculty/Stakeholders	5.	Dissemination of results
6. Document the Use of Results for Improvement	6a.	Improvement of programs regarding student learning and development
	6b.	Improvement of assessment process

2009-2010 Procedures. The rating of the 2009-2010 APTs occurred in two stages. In stage one, rater teams evaluated different sets of APTs (i.e., raters nested within teams). Rating stage two incorporated a more robust design in which all raters evaluated the same set of APTs (i.e., no nesting). For the purposes of this research, the focus is on stage two in which the fully-crossed design was used to calculate variance components and dependability estimates.

For stage one, 12 graduate students with expertise in assessment and measurement techniques participated in one full day of rater training. At the conclusion of the training, six sets of two-person rater teams were randomly assigned samples of approximately 16 APTs to evaluate using the rubric. In generalizability theory terms, this process is a (program by raters)-nested-within-teams ((p x r) : team) design.

Programs without assessment results –those receiving a "1" for element 4a of the rubric - were not considered for stage two. Because several APT rubric elements cannot be rated without results, programs lacking assessment results were not eligible for the fully-crossed design. These reports provided limited information (i.e., all raters would have rated these programs a one on several elements), which in turn would have artificially inflated rater reliability between elements. After programs without results were removed, 85 programs were eligible for the fully-crossed design.

In a fully-crossed design, all objects of measurement (i.e., academic programs) are rated by all raters (i.e., all academic programs rated by 12 raters). However, one benefit of G-theory is that stable estimates of dependability can be gleaned from relatively small samples of the object of measurement. Therefore, a subset of programs was used in the study. To choose this subset, the ratings from each rater team during stage one were aggregated and all 85 programs were rank-ordered based on their total

average element score on the APT. Average scores could range from one to four. Although comparing programs between rater teams is not entirely appropriate—the scores of each program cannot be totally distinguished from the personal subjectivity of each rater team—the program comparisons provided a way to select APTs for stage two that helped ensure variability in the quality of the reports.

When establishing stable estimates of dependability it is important that there is variability among the sample. To that end, during stage two a stratified random sampling process was conducted to select the program APTs for the fully-crossed design. To ensure that the subset of programs represented a wide range of scores on the APT, the 85 eligible program APTs were split into sextiles, with approximately 14 program APTs per group. From each sextile, two programs were chosen at random, resulting in 12 program APTs to be used for the fully-crossed design. This sampling method helped to ensure that the programs chosen would represent the variability of the population of APTs, and would not artificially deflate the dependability estimates.

Nine graduate student raters independently provided ratings on the 14 APT rubric elements for the 12 chosen program APTs. All but one of the raters had participated in the stage one rating process and all raters were provided a one-hour condensed training prior to scoring the 12 programs. The one rater who had not participated in stage one of the rating process received the complete rater training prior to beginning the stage two rating process. Because the stage two APTs were also used during the stage one rating process, raters had previous exposure to at least one of the 12 programs. Raters were told to provide scores on the 12 programs as if they had not seen them before, in order to limit bias. Ratings for stage two were conducted over a one-week period.

2010-2011 Procedures. The main difference between the two years is the inclusion of faculty raters during the 2010-2011 study. To that end, eight faculty raters were recruited—one from each college of the university to ensure a breadth of representation across disciplines—to participate in the 2010-2011 rating process. Each faculty member was paid a stipend for the two weeks that he/she was involved in this study. Similar to the students in the 2009-2010 study, faculty members participated in rater training to orient and calibrate them on the use of the rubric. However, the faculty rater training consisted of an additional day in which they were presented with foundational information about the six stages of assessment. This training was designed to educate faculty raters with very limited assessment knowledge.

Additionally, eight graduate students also participated in the second day of training alongside the faculty members. During this day, all participants calibrated their ratings on example assessment reports. The second day of rater training closely mirrored the training provided to the graduate students in 2009-2010. After the training, faculty members and graduate students were paired together—one faculty member with one graduate student—and each team was randomly assigned a group of approximately 15 APTs to rate.

Over the span of seven days, each rater team independently rated their assigned APTs, meeting at regular intervals to discuss discrepancies in their ratings that varied by more than a point. Again, this process of adjudicating discrepant ratings was the same process used by graduate students during stage one of the 2009-2010 rating process.

Following the initial rating of the 2010-2011 APTs, each of the eight faculty raters were then tasked with independently rating the same 12 APTs that graduate students rated during the 2009-2010 fully-crossed design. By using the same APTs, the

results from the faculty ratings were directly compared to those of the graduate student ratings. Faculty rated these 12 APTs over three days. It should be noted that in both the 2009-2010 and 2010-2011 fully-crossed studies, there were no missing data; each rater assigned scores to all elements for each APT assigned with no exceptions.

Identical to the procedure used in the 2009-2010 study, generalizability theory was used to estimate the variance components for all relevant sources of variance in the universe of admissible observations (UAO). In the UAO, APT scores represented the object of measurement. Raters (graduate student and faculty) were considered random facets (i.e., the universe of possible raters was assumed to be larger than the sample used in either study). The 12 program APTs were fully crossed with eight raters (p x r). The 14 elements were also a source of systematic variance, but this facet was considered fixed for two reasons. First, a facet is fixed if there are no additional elements available in the UAO from which to sample. The APT rubric was developed to evaluate the elements that best represent the assessment process, and therefore, it is assumed that the entire construct is represented in some way on the rubric. Thus, there is no strong theoretical basis for switching the 14 rubric elements with another sample of potential elements (i.e., the elements are not exchangeable; Shavelson & Webb, 1991).

Second, the Element facet is fixed because in this meta-assessment the variance associated with differing element means is not error. Unlike raters, where consistency between graduate students or faculty members is the goal, programs are expected to score differently among the elements, and these trends are relatively predictable. The vast majority of programs, for example, have student-centered objectives and therefore the average score on this element is quite high (i.e., above three). Other elements such as the presence of reliability estimates for example, are not as well-understood by programs,

and this information is only presented in a handful of reports where faculty have advanced knowledge of assessment or have sought out consultation. Therefore, it is expected that programs as a whole will score lower on this element than on others. In many applications of G-theory regarding absolute decisions, one would desire to have equally difficult items or elements in order to minimize the error variability in this facet. Because in this study, the element difficulties are expected to vary, if this facet were considered random, then error would incorrectly be introduced into the dependability estimates, artificially lowering them.

To address research question one, G- and D-studies were completed on the faculty ratings to estimate the relative and absolute dependability of the current universe of generalization (i.e., two raters scoring all 14 elements for all 12 programs). Guidelines for acceptable values of phi- and G-coefficients have not been established in the literature. However, given the similarities between generalizability theory and classical test theory, one may consider the guidelines for the internal consistency estimate of Cronbach's alpha to be a suitable alternative. In fact, the G-coefficient will be identical to Cronbach's alpha in one-facet designs. George and Mallery (2003) suggest certain rules of thumb for interpreting Cronbach's alpha, which can be found in Table 6. Given these rules of thumb, a value of .80 was used as a general guideline for good estimates of relative dependability. While the .80 cutoff for a G-coefficient and Cronbach's alpha is roughly analogous, there is really no good cutoff analog for a phi-coefficient. The reader may recall that phi-coefficients incorporate more sources of error and thus are almost always lower than G-coefficients. With no other alternative, the .80 will still be used as a cutoff for the phi-coefficient. The reader may interpret this cutoff as a more rigorous benchmark.

To address research question two, descriptive analyses were used to compare the dependability estimates of faculty raters to the graduate student dependability estimates obtained in the 2009-2010 study. Research question three was addressed by comparing the average score across all 12 programs and raters between graduate students from 2009-2010 and faculty members from 2010-2011. First, program mean comparisons between graduate students and faculty raters were examined to identify potential patterns in harshness or leniency. However, mean comparisons alone yield limited information about the sources of variability that contributed to any mean score differences between graduate students and faculty members. Thus, to explore the variation in program means between rater groups further and to check for interactions, an additional G- and D-study was performed. In this study, ratings from both the 2009-2010 and 2010-2011 studies were combined. To account for the different rater groups (Faculty and Graduate Student), an additional facet—rater type (t)—was included in the design. By adding the rater type facet, the design became unbalanced (nine graduate students and eight faculty members). Furthermore, raters were now nested in rater type. That is, a rater can only be a graduate student or a faculty member, not both. The resulting unbalanced design then, was no longer fully crossed, but instead became partially nested (p x i x (r:t)).

Finally, research question four will be addressed by calculating the standard errors of measurement of each element score for the faculty ratings. These values will be compared to the standard errors calculated from the graduate student ratings in the 2009-2010 study to determine whether certain elements led to inconsistencies among both rater groups. A comparison across rater type of these standard errors will not only help identify elements that are interpreted differently between graduate students and faculty members, but it may lead to interventions in rater training or changes to the rubric

designed to limit future measurement error. The analyses for research questions one, two, and four were conducted using the software package GENOVA (Crick & Brennan, 2001). The analysis for research question three was conducted with urGENOVA software, which can handle unbalanced designs (Brennan, 2001).

Table 6

Rules of Thumb for Estimates of Cronbach's Alpha

Value	*Label*
>.90	Excellent
>.80	Good
>.70	Acceptable
>.60	Questionable
>.50	Poor
<.50	Unacceptable

Chapter 4

Results

Dependability of Faculty Ratings

A G-study of the faculty ratings yielded the variance components in Table 7. The G-study variance components reflect the systematic variance for a model that treats the Element facet as random. In the D-studies, however, the Element facet is treated as fixed; the variance associated with the main effect of elements and subsequent interactions (e.g., Program x Element, Rater x Element, Program x Rater x Element plus Random Error) are not counted as systematic error because fixed facets are not generalized to a larger universe. Therefore the variability is not error, but rather, it is subsumed by the object of measurement variance component (Programs) in the D-studies. However, to better understand the sources of variability in the APT rating design, the variance components containing the Element facet were estimated in the G-studies.

The Program variance component ($\hat{\sigma}^2_p$) in the G-study was 0.3195, indicating that approximately 31.2 percent of the total variance in the ratings was due to differences in ability between the programs. In comparison, the Rater facet (i.e., the amount of variability due to differences in rater stringency; $\hat{\sigma}^2_r$ = .0160) accounted for 1.6 percent of the total variance. The Program x Rater variance component ($\hat{\sigma}^2_{pr}$ = .0542) indicated that 5.3 percent of the total variance was because of differences in the relative rank order of programs by raters.

Table 7

2010-2011 APT Ratings Using the Fully Crossed Design: Contribution of each Facet to Score Variance

		G Study[a]		D Study[b]	
Source of Variation	$\hat{\sigma}^2$	*nr* = 1 *ni* = 1	% Total Variance	*nr* = 2 *ni* = 14	% Total Variance
Program (*p*)	$\hat{\sigma}^2_p$	0.3195	31.2	0.3315	88.0
Rater (*r*)	$\hat{\sigma}^2_r$	0.0160	1.6	0.0085	2.2
Element (*i*)	$\hat{\sigma}^2_i$	0.1763	17.2	---	---
pr	$\hat{\sigma}^2_{pr}$	0.0542	5.3	0.0369	9.8
pi	$\hat{\sigma}^2_{pi}$	0.1678	16.4	---	---
ri	$\hat{\sigma}^2_{ri}$	0.0139	1.3	---	---
pri,e	$\hat{\sigma}^2_{pri,e}$	0.2746	26.8	---	---
$\hat{\sigma}^2_{REL}$				0.0369	
$\hat{\sigma}^2_{ABS}$				0.0454	
$\hat{\rho}^2$				0.90	
$\hat{\Phi}$				0.88	

Note. [a]Variance components if the Element facet (i) is random.
[b]Element facet (i) is treated as fixed.

The remaining variance components contain the Element facet. The Element facet ($\hat{\sigma}^2_i$ = .1763), when treated as a random facet, contributed 17.2 percent to the total variability in the ratings. This is the amount of variability between the element averages across all raters and programs. Similarly, the Program x Element facet ($\hat{\sigma}^2_{pi}$ = .1678) accounted for 16.4 percent of total variance, indicating that, across all raters, the relative difficulty of elements differed depending on the program. The Rater x Element facet ($\hat{\sigma}^2_{ri}$ = .0139), which illustrates the relative harshness between raters on particular elements, accounted for only 1.3 of the total variance. The final variance component in

the G-study ($\hat{\sigma}^2_{pri,e}$ = .2746), accounted for 26.8 percent of the total variance. This variance component is a mixture of the variance attributable to the Program x Rater x Element interaction and the random error still remaining in the model. They are combined because the random error cannot be disentangled from the *pri* variance component.

Results from the D-study, in which the Element facet is fixed, illustrate the effects of using a particular rating design. By fixing the Element facet, the Program variance component increased from .3195 in the G-study to .3315 in a two-rater by 14-element D-study. Equation 5 illustrates how the Program variance component in the D-study (*p**) was calculated. Similarly, the Program x Rater (*pr**) variance component subsumed the variability due to the *pri,e* interaction averaged over the levels of the facets (see Equation 6).

$$\hat{\sigma}^2_{p^*} = \hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pi}}{n_i} \quad [5]$$

$$\hat{\sigma}^2_{pr^*} = \frac{\hat{\sigma}^2_{pr}}{n_r} + \frac{\hat{\sigma}^2_{pri,e}}{n_r n_i} \quad [6]$$

The results from the D-study (see Table 7) illustrate the dependability under a set of conditions specified by the researcher. Here, the variance components, relative ($\hat{\sigma}^2_{REL}$) and absolute ($\hat{\sigma}^2_{ABS}$) variability coefficients, the G-coefficient ($\hat{\rho}^2$), and the phi-coefficient ($\hat{\Phi}$) represent the dependability of the ratings under the conditions of measurement used to rate APTs. That is, the estimates reflect the dependability of two raters scoring all fourteen elements. The relative variability ($\hat{\sigma}^2_{REL}$) is .0363. This value is comprised of the variance components affecting the rank ordering of programs (e.g.,

pr). The absolute variability ($\hat{\sigma}^2_{ABS}$), which consists of the variance components that impact the variance of scores relative to the APT rubric anchors (e.g., r, pr), is .0454. The G-coefficient ($\hat{\rho}^2$) for the two-rater by 14-element design is .90 and the phi coefficient ($\hat{\Phi}$) is .88.

Comparison of Results from Fully-Crossed Designs Involving Graduate Students and Faculty

Table 8 provides a comparison of the variance components and dependability estimates calculated for both the graduate student ratings (2009-2010) and the faculty ratings (2010-2011). According to the G-study, the variance component for Programs was higher for the faculty ratings than for the graduate students' (Faculty $\hat{\sigma}^2_p$ = .3195; Graduate Student $\hat{\sigma}^2_p$ =.2171). This finding indicates that across all raters and elements, there was greater variability among program means when faculty rated as opposed to graduate students. The results illustrate that differences in element difficulty account for a large portion of the total variance for both faculty and for graduate student raters (Faculty $\hat{\sigma}^2_i$ = .1763; Graduate Students $\hat{\sigma}^2_i$ = .2540). The Rater facet, although relatively small for both groups of raters, was twice as large for faculty as for graduate students (Faculty $\hat{\sigma}^2_r$ =.0160; Graduate Student $\hat{\sigma}^2_r$ = .0055). This finding reveals that faculty raters, across all programs and elements, were about twice as variable in their ratings as were graduate students.

The facet interactions also yielded some interesting comparisons. The Program x Rater facet was twice as great for faculty as it was for graduate students (Faculty $\hat{\sigma}^2_{pr}$ = .0542; Graduate Students $\hat{\sigma}^2_{pr}$ = .0228), revealing that the rank order of programs varied

more between faculty raters than it did for graduate student raters. The Program x Element variance components were comparable between the two groups (Faculty $\hat{\sigma}^2_{pi}$ = .1678; Graduate Students $\hat{\sigma}^2_{pi}$ = .1961), and showed that across raters, the rank order of programs differed depending on the element in question. The Rater x Element component was quite similar between the two rater groups (Faculty $\hat{\sigma}^2_{pi}$ = .0139; Graduate Students $\hat{\sigma}^2_{pi}$ = .0125). The final variance component, $\hat{\sigma}^2_{pri,e}$, was larger for faculty (.2746) than for graduate students (.2088), but because of this component's complexity, it is unknown whether the additional variance among faculty was due to random error or the specific Program x Rater x Element interaction

Results from the two-rater, 14-element D-study for faculty were compared to the results from the equivalent D-study for graduate students (See Table 8). Both the relative and absolute variances for faculty were twice that for the graduate students (Faculty $\hat{\sigma}^2_{REL}$ = .0363, Graduate Student $\hat{\sigma}^2_{REL}$ =.0189; Faculty $\hat{\sigma}^2_{ABS}$ = .0454, Graduate Student $\hat{\sigma}^2_{ABS}$ = .0221) revealing that faculty raters were more varied in their rank ordering of programs as well as in their ability to rate programs consistently against the APT rubric. When comparing dependability estimates between the two groups of raters, the G- and phi-coefficients for faculty are slightly lower than those of the graduate students ($\hat{\rho}^2_{\text{Faculty}}$ = .90, $\hat{\Phi}_{Faculty}$ = .88; $\hat{\rho}^2_{\text{GS}}$ = .92, $\hat{\Phi}_{GS}$ = .91).

Table 8

2009-2010 and 2010-2011 G- and D-study Results: Comparison of Graduate Students and Faculty Members

Source of Variation	$\hat{\sigma}^2$	G Study[a] *nr* = 1 *ni* = 1 Graduate Students	G Study[a] *nr* = 1 *ni* = 1 Faculty Members	D Studies[b] *nr* = 2 *ni* = 14 Graduate Students	D Studies[b] *nr* = 2 *ni* = 14 Faculty Members
Program (*p*)	$\hat{\sigma}^2_p$	0.2171	0.3195	0.2310	0.3315
Rater (*r*)	$\hat{\sigma}^2_r$	0.0055	0.0160	0.0032	0.0085
Element (*i*)	$\hat{\sigma}^2_i$	0.2540	0.1763	---	---
pr	$\hat{\sigma}^2_{pr}$	0.0228	0.0542	0.0188	0.0369
pi	$\hat{\sigma}^2_{pi}$	0.1961	0.1678	---	---
ri	$\hat{\sigma}^2_{ri}$	0.0125	0.0139	---	---
pri,e	$\hat{\sigma}^2_{pri,e}$	0.2088	0.2746	---	---
	$\hat{\sigma}^2_{REL}$			0.0189	0.0369
	$\hat{\sigma}^2_{ABS}$			0.0221	0.0454
	$\hat{\rho}^2$			0.92	0.90
	$\hat{\Phi}$			0.91	0.88

Note. [a]Variance components if the Element facet (i) is random. [b]Element facet (i) is treated as fixed.

Rater Stringency

The mean scores for each of the twelve programs across all raters were compared to determine the relative leniency and harshness of the raters. These average scores, including the total average score across all programs, are located in Table 9. Results are presented in order of the average element score assigned by graduate students. The program rated the highest by graduate students is first (Program 10) with the program receiving the lowest average score from graduate students listed last (Program 3). The total average score for each rater group was very similar (M_{GS} = 2.86, $M_{Faculty}$ = 2.91). Program mean comparisons are also found in Table 9. Both rater groups scored program three the lowest (M_{Fac} = 1.77; M_{GS} = 1.96) and program 10 the highest (M_{Fac} = 3.63; M_{GS} = 3.49). Differences in mean scores between rater groups ranged from .02 points (Program Four) to .28 points (Program Eight), with a median difference in scores of .11 points. Average faculty ratings were higher than graduate students for eight of the 12 programs.

Table 10 contains the results of the G- and D-studies in which results from both fully crossed designs were combined. In this design, the Rater Type facet (*t*) and the nesting of raters within rater type (*r:t*) are both modeled along with their subsequent interactions with other facets. As with the previous designs, to calculate the G-study variance components, all facets were treated as random. Results of the G-study indicated that the program variance component accounts for approximately 26.3 percent of the total variance in the combined design ($\hat{\sigma}^2_p$ = .2613). In comparison, program variance accounted for 23.6 and 31.2 percent of the total variances in the 2009-2010 and 2010-2011 fully crossed designs respectively. Rater Type (*t*) accounted for a negligible

percentage of total variance, whereas the Element facet (*i*) explained 21 percent of the total variability, consistent with findings from the 2009-2010 and 2010-2011 studies. Raters nested within rater type explained 3.8 percent of total variance.

Table 9

Program Average Element Scores for Graduate Students and Faculty Raters According to Graduate Student Ratings

Program	2009-2010 Ratings (Graduate Students)	2010-2011 Ratings (Faculty)	Score Difference[a]
10	3.49	3.63	-0.14
7	3.40	3.31	0.09
4	3.18	3.2	-0.02
11	3.15	3.23	-0.08
9	3.13	3.38	-0.25
5	3.03	3.15	-0.12
8	2.94	3.22	-0.28
1	2.85	2.92	-0.07
2	2.64	2.74	-0.10
6	2.61	2.55	0.06
12	2.03	1.88	0.15
3	1.96	1.77	0.19
Total	2.86	2.91	-.05

Note. [a]Score Difference indicates the relative stringency of faculty relative to graduate students. Negative values mean that faculty, on average, were more lenient for a given program. Positive values indicate that graduate students were more lenient on average.

The Program x Rater Type (*pt)* variance component was .0041. This component, which is interpreted as the relative differences in program rank order between rater types, accounted for .4 percent of the total variance. The Program x Rater Type x Element (*pti*) facet accounted for 17.9 percent of the total variance whereas the *pi, ti, (r:t) x p,* and *(r:t) x i* facets accounted for 6.5 percent of the total variance combined. Finally, the *pi x (r:t)* facet, which includes random error, accounted for the remaining 24.1 percent of variance in the model.

The D-study in Table 10 reflects a design in which the Rater Type facet is random and Element facet is fixed. The g- and phi-coefficients indicate that this model produces relative dependability ($\hat{\rho}^2$) of .90 and absolute dependability ($\hat{\Phi}$) of .89. The proportion of total variance attributable to Programs increases from 26.3 percent to 89.2 percent when the Element facet is fixed and the Rater Type facet is random.

Table 10

Variance Components and Dependability Estimates for Combined 2009-2010 and 2010-2011 Ratings

		G Study[a]		D-Study[b]	
Source of Variation	$\hat{\sigma}^2$	*nr:t* = 1 *nt* = 1 *ni* = 1	% Total Variance	*nr:t* = 1 *nt* = 2 *ni* = 14	% Total Variance
Program (*p*)	$\hat{\sigma}^2_p$	0.2613	26.3	0.2740	89.2
Rater Type (*t*)	$\hat{\sigma}^2_t$	0.0000	0.0	0.0004	0.2
Element (*i)*	$\hat{\sigma}^2_i$	0.2088	21.0	---	---
Raters:Type *(r:t)*	$\hat{\sigma}^2_{r:t}$	0.0375	3.8	0.0032	1.0
pt	$\hat{\sigma}^2_{pt}$	0.0041	0.4	0.0022	0.7
pi	$\hat{\sigma}^2_{pt}$	0.1782	17.9	---	---
(r:t) x p	$\hat{\sigma}^2_{p(r:t)}$	0.0375	3.8	0.0273	8.9
ti	$\hat{\sigma}^2_{ti}$	0.0087	0.9	---	---
(r:t) x i	$\hat{\sigma}^2_{i(r:t)}$	0.0132	1.3	---	---
pti	$\hat{\sigma}^2_{pti}$	0.0049	0.5	---	---
pi(r:t),e	$\hat{\sigma}^2_{pi(r:t),e}$	0.2395	24.1	---	---
	$\hat{\sigma}^2_{REL}$			0.0295	
	$\hat{\sigma}^2_{ABS}$			0.0332	
	$\hat{\rho}^2$			0.90	
	$\hat{\Phi}$			0.89	

Note. [a]Variance components if the Element and Rater Type facets are random. [b]Element facet is fixed and Rater Type facet is treated as random

Element Means and Standard Errors of Measurement

The standard errors of measurement provide an indication of the precision of each element's ratings; they are the average distances that ratings fall from the elements' universe scores. Element means and absolute standard errors of measurement were analyzed for both faculty and graduate student ratings (see Table 11). The absolute standard errors of measurement represent the precision of element scores relative to the rubric anchors, and are therefore of greater importance to answer this research question. To calculate the absolute standard errors of measurement, G-studies were run on ratings for each element separately. Because each G-study was conducted to determine the rater variance components associated with a particular element, the Rater facet was the only source of systematic error modeled in each study. Thus, 14 one-facet G-studies were conducted on faculty ratings from 2010-2011 and 14 one-facet G-studies were conducted on graduate student ratings from 2009-2010. The absolute standard error of measurement was then derived for each element by taking the square root of the absolute variance component from a D-study design using two raters.

Graduate students rated programs slightly lower than faculty and had a smaller overall standard error (M_{Fac} = 2.91 SE_{Fac} = .21, M_{GS} = 2.86 SE_{GS} = .15). For both groups, Element 2 (Curriculum Map) had the highest average score (M_{Fac} = 3.67 SE_{Fac} = .36, M_{GS} = 3.75 SE_{GS} = .24) and Element 3e (Additional Validity Evidence) was rated the lowest (M_{Fac} = 2.04 SE_{Fac} = .42, M_{GS} = 1.73 SE_{GS} = .23).

Table 12 provides the standard errors for both faculty and graduate students for all fourteen elements. For graduate students, Element 3e (Additional Validity Evidence) had the smallest standard error (SE=.23) whereas the element with the smallest standard error

for faculty was Element 3a (Measure-to-Objective Match; SE=.32). Graduate students demonstrated the most variability in rating Element 3c (Specification of Desired Results; SE = .47), and the element with the greatest standard error for faculty was Element 1b (Orientation; SE=.51).

Table 11

Rank Order of Elements by Mean Score: Comparison of Graduate Students to Faculty

2009-2010 Ratings (Graduate Students)			2010-2011 Ratings (Faculty)		
Element	*M*	SE_M	*Element*	*M*	SE_M
3e	1.73	0.23	3e	2.04	0.42
3c	2.41	0.47	3c	2.26	0.50
6b	2.57	0.32	6b	2.56	0.43
3d	2.62	0.34	4c	2.75	0.36
4c	2.63	0.39	3d	2.76	0.40
6a	2.67	0.44	6a	2.77	0.45
3a	2.81	0.31	5	2.97	0.41
4a	2.89	0.35	4a	3.01	0.39
1a	2.93	0.32	4b	3.06	0.48
5	2.95	0.33	3a	3.09	0.32
4b	3.10	0.45	1a	3.14	0.41
3b	3.40	0.42	3b	3.29	0.45
1b	3.69	0.23	1b	3.46	0.51
2	3.75	0.24	2	3.67	0.36
Total	2.86	0.15	Total	2.91	0.21

Note. The standard errors of measurement for the Total scores were calculated using the phi coefficients from the two-rater x fourteen-element design

Table 12

Rank Order of Elements by Standard Errors: Comparison of Graduate Students to Faculty

	2009-2010 Ratings (Graduate Students)			2010-2011 Ratings (Faculty)	
Element	*M*	SE_M	*Element*	*M*	SE_M
3e	1.73	0.23	3a	3.09	0.32
1b	3.69	0.23	4c	2.75	0.36
2	3.75	0.24	2	3.67	0.36
3a	2.81	0.31	4a	3.01	0.39
6b	2.57	0.32	3d	2.76	0.40
1a	2.93	0.32	5	2.97	0.41
5	2.95	0.33	1a	3.14	0.41
3d	2.62	0.34	3e	2.04	0.42
4a	2.89	0.35	6b	2.56	0.43
4c	2.63	0.39	6a	2.77	0.45
3b	3.40	0.42	3b	3.29	0.45
6a	2.67	0.44	4b	3.06	0.48
4b	3.10	0.45	3c	2.26	0.50
3c	2.41	0.47	1b	3.46	0.51
Total	2.86	0.15	Total	2.91	0.21

Note. The standard errors of measurement for the Total scores were calculated using the phi coefficients from the two-rater x fourteen-element design

Chapter Five

Discussion

One of the major research questions posed in this study was whether or not faculty members could dependably rate using a meta-assessment rubric. Results showed that two faculty could indeed rate programs dependably across all fourteen elements ($\hat{\rho}^2$ = .90, $\hat{\Phi}$ =.88). Using this two-rater design, faculty members produced ratings that were well above the .80 benchmark for adequate dependability. The high G-coefficient indicates that faculty rank ordered programs consistently. The high phi-coefficient (the estimate of absolute dependability) provides evidence that, in addition to rank ordering programs consistently, faculty also tended to rate programs similarly with respect to the rubric anchors. Therefore, faculty raters not only rank ordered programs similarly, but they tended to produce scores that were consistent with their fellow raters as well.

Given that G-theory provides two estimates of dependability, one may question whether the g- or the phi-coefficient should be scrutinized more within the context of this study. If one were only interested in the rank order of programs relative to each other, then the g-coefficient provides ample information for that purpose. Regarding the APT, however, programs are concerned with how their assessment scores compare to the rubric's behavioral anchors (i.e., absolute dependability). That is, a program does not receive feedback about its assessment practice relative to other academic programs, but rather, the rubric scores are used to provide programs with information about the quality of its assessment as defined by the rubric. Therefore conceptually, the information one can gather from the phi-coefficient is more relevant for this research than the information provided by the g-coefficient.

The dependability of faculty ratings was comparable to that of graduate student ratings. The faculty rater and program by rater facets were larger than those of graduate students (r_{fac} = .0160, r_{GS} =.0055; pr_{fac} = .0542, pr_{GS} = .0228). However, these variance components are quite small, indicating relatively unsubstantial differences between the two rater types. For example, consider the range of rater means across all programs and elements. To calculate the range, the square root of the variance component is computed (Faculty $\hat{\sigma}_r$ = .13). This value is the standard deviation of the average faculty rating across all programs and elements. Assuming a normal distribution, four standard deviations will encompass approximately 95% of rater scores. Thus, the majority of rater average scores will range .52 points on the four point rubric scale, roughly half a point on the four point scale. The standard deviation of graduate students, in comparison, is .07 and a range of 95% of graduate student average ratings is .28 points. From a practical interpretation of these ranges, a half point difference on the ratings is not very substantial, nor is a range of .28 points. In essence, a faculty rater at the low end of the range might have an average score across all programs of 2.5 whereas a faculty rater at the other end of the range may have an average of 3.0. The range of average ratings for graduate students is even tighter, with a potential difference between a harsh graduate student rater and a lenient graduate student rater fluctuating a quarter point on the four-point scale.

The differences in the Rater and the Rater x Program variance components between faculty and graduate students likely contributed to small differences in the dependability estimates. Because of the larger variance components related to raters, the faculty G- and phi-coefficients for a two-rater, 14-element design were slightly below those of the graduate students ($\hat{\rho}^2_{Fac}$ = .90, $\hat{\rho}^2_{GS}$ = .92; $\hat{\Phi}_{Fac}$ = .88, $\hat{\Phi}_{GS}$ = .91),

indicating that using two faculty to rate the 14 elements would yield slightly less precise scores than using graduate student raters under similar conditions. To account for the larger rater variance components among faculty, three faculty raters would need to score each APT. This alternate design would produce a G-coefficient of .93 and a phi-coefficient of .92. However, for such a small increase in dependability, the additional investment in time and resources for this alternative design hardly seems appropriate, especially when one considers the similar dependability estimates for faculty and graduate students under the two-rater design.

The dependability estimates of faculty and graduate students provide valuable information about the consistency with which teams of faculty raters score assessment reports relative to the consistency of graduate student rater teams. However, one may question whether faculty raters as a group were consistently harsher or more lenient than graduate students. Thus, even though faculty raters were consistent with one another, their overall program scores may be substantially lower—or higher—than the scores granted by graduate students. If the two types of raters are interpreting the rubric differently, this potential finding would call to question the interchangeability of raters. In other words, could a team comprised of one faculty and one graduate student rater produce scores with the same consistency as rater teams of faculty or graduate students only?

The differences in program means between the rater groups provide a starting place for exploring rater stringency. The average score across all the programs was very similar between faculty (M=2.91, SD=.58) and graduate students (M=2.86, SD=.54). Further, the individual program means were quite close between the two rater types.

Faculty raters were consistently more lenient than graduate students, but their relative leniency was hardly substantial as the mean difference in total average score between the rater types was .05 points. Further, faculty tended to score the programs with weaker assessment reports harsher than graduate students. Likewise, faculty also rated programs with stronger reports higher than graduate students. The more extreme scores at either ends of the rating spectrum were likely the main contributing factor to the increased Program variance among faculty.

It is evident that faculty were generally more lenient in ratings, though they rated programs at the low end of the assessment spectrum harsher than graduate students. Comparing program means is a quick and easy method for determining relative rater stringency; however, the results of the combined faculty and graduate student analysis allow for a richer interpretation of the specific sources of variance contributing to rating differences between groups. This point warrants further discussion. When ratings from both rater groups were combined and analyzed together, the variance components for Programs, the Element facet, the Program x Element interaction, and the Program x Element x Raters nested within Rater Type interaction accounted for the largest proportions of total variance ($\hat{\sigma}^2_p$ = 27.3 percent; $\hat{\sigma}^2_i$ = 21.0 percent; $\hat{\sigma}^2_{pi}$ = 17.9 percent; $\hat{\sigma}^2_{pi(r:t)}$ = 24.1 percent). These findings were very comparable to the results from the separate 2009-2010 and the 2010-2011 studies. Additionally, the variance components involving the Rater Type and Raters nested within Rater Type facets were quite small, indicating that overall, graduate students did not vary much from faculty, nor did the individual raters within each type vary much from each other.

The D-study results further demonstrate the consistency between raters in a design where the Rater Type facet is treated as random. Because the Rater Type facet is treated as random, the universe of generalization is assumed to include additional kinds of raters above and beyond graduate students and faculty members. Other types of raters may include staff members or administrators, all of whom could rate assessment reports if provided the same training as graduate students or faculty. Further, if Rater Type is a random facet, then the combination of raters within teams is considered more versatile than if the facet is fixed. That is, the dependability estimates in Table 10 represent the consistency of a two rater team in which raters could consist of two graduate students, two faculty members, or one faculty member and one graduate student. Thus, with a random Rater Type facet, one can determine, to some degree, the interchangeability of rater types. The results show that two faculty members could rate programs as dependably as two graduate students or a combination of graduate students and faculty members. If the Rater Type facet had been fixed, rater teams would be assumed only to include one faculty member and one graduate student. Therefore, treating the Rater Type facet as random extends the interpretations one can make about the ratings and the consistency of different combinations of raters.

Ratings between programs were shown to be similar between graduate students and faculty members, but what about the precision with which the elements were rated by each rater type? The elements' standard errors provide an estimation of rating precision taking into account the dependability of the scores for individual elements. Elements with larger standard errors indicate areas of the rubric where raters tended to be more variable in their scores. By examining the standard errors, rater inconsistencies within

elements can be identified. Future trainings can then be adapted to focus on those elements that led to the most inconsistencies among raters. Further, comparisons of element standard errors by rater type may uncover elements that lead to greater inconsistencies in one type of rater than the other. Locating these inconsistencies may have implications for how graduate students are trained versus faculty members.

Graduate students tended to rate the elements overall with slightly better precision than faculty. The overall standard error across all elements was .15 for graduate students compared to a standard error of .21 for faculty. Graduate student ratings were the most precise for Element 3e (Additional Validity Evidence; SE=.23), 1b (Orientation; SE=.23), and 2 (Course/Learning Experiences; SE=.24). Given that Element 3e was rated the most difficult (M=1.73) and Elements 1b (M=3.69) and 2 (M=3.75) were rated the easiest by graduate students, it is possible that certain ceiling effects limited the variability in scores, leading to greater consistency among the ratings. However, even though faculty also rated Element 1b high (M=3.46), the element had the largest standard error (SE = .51). It is evident that certain faculty were likely misinterpreting how to rate this particular element, whereas certain faculty and the majority of graduate students were interpreting the element similarly. Thus, future training should focus on this element, particularly among faculty members, so that inconsistent interpretations are minimized.

The standard errors also illustrate the appropriateness of using faculty and graduate students as raters. For example, graduate students had a substantially smaller standard error for Element 3e (Additional Validity Evidence) than did faculty (SE_{Fac} = .42; SE_{GS} = .23). To rate this element effectively, the rater should have specific knowledge of reliability estimates and the various types of validity. The graduate student

raters receive this training through coursework and assistantships. Faculty, depending on their prior assessment experience and field of expertise, may have little to no capacity to identify appropriate evidence of good reliability and validity. This fact likely contributed to the faculty's larger standard error of measurement associated with this element.

Furthermore, faculty consistently rated programs higher than graduate students on Element 3e ($M_{Fac} = 2.04$; $M_{GS} = 1.73$). This finding suggests that faculty were more willing to rate programs higher on this element, possibly giving credit for programs simply mentioning reliability or validity regardless of the accuracy of the program's claims. Faculty may also give their colleagues "the benefit of the doubt" with regards to this element because the raters themselves do not understand what is being reported, nor do they fully understand the rationale for including such information in an assessment report. Rater training should then be focused more on helping faculty recognize the importance of validity evidence and how to identify poor reporting. However, both groups rated programs the lowest on Element 3e, which suggests that although faculty do not rate this element as precisely as graduate students, nor are they as harsh as student raters, the two rater groups can still form consistent conclusions about the relative quality of validity evidence provided by programs.

Elements 3c (Specification of Desired Results), 4b (History of Results), 6a (Program Improvements), and 3b (Types of Measures) produced large standard errors for both rater types, with values ranging from .42 to .47 for graduate students and .45 to .50 for faculty. Because these elements tended to produce inconsistencies among both types of raters, more attention should be paid to calibrating these ratings in future trainings. The imprecision of ratings for Element 3c is not entirely surprising. This element

requires programs to specify and justify expected results that help them determine the degree to which students met the objective (e.g., Students will average an 80% on the assessment exam). In addition to specifying their expected results, to receive an exemplary score on this element programs also have to justify why they have chosen their specific targets. Thus, the degree to which programs justify, and even specify, their desired results varies in both detail and quality of response, leading to more subjective interpretations by raters and greater imprecision in the scores. Additional training can cover multiple examples of responses raters might encounter in order to improve objective interpretations.

Although the ratings for elements such as 3c have potential to be more subjective, the fact that some elements were rated inconsistently is cause for some concern. For example, both faculty and graduate student raters scored Element 4b inconsistently. This element addresses whether or not programs provide multiple years of assessment results in their reports. With many program reports, this element is fairly straightforward. That is, programs provide tables of results with multiple columns, one for each year of data. However, some programs report data by semesters, whereas other programs provide multiple years of data for some of their assessment measures but not for others. Raters may interpret these cases differently, leading to overall inconsistencies in the element score. In future training sessions, raters should be provided with specific instructions of special cases they may find during the rating process and instructions on how to rate these reports.

Although the precision of ratings for elements 3c and 4b could improve with additional training, the strategies for increasing the precision may differ depending on the

subjective nature of the particular element. For elements such as 3c where raters have to interpret the contents of the report to make a scoring judgment, the training should allow for more discussion time and multiple examples of cases that raters may encounter. Elements such as 4b, however, may require more specific guidance from rating facilitators who can provide directions of how to rate a program based on particular situations.

Limitations

The current research presented some limitations. For one, because graduate students and faculty completed their respective fully-crossed studies one year apart, the combined ratings incorporated scores over two different occasions. Therefore, a facet that accounts for variations in scores resulting from the time difference of the studies may have been of importance. The Rater Type facet represents the same variability that the occasion facet includes. Because the 2009-2010 study consisted of all graduate students and the 2010-2011 study was all faculty raters, the Rater Type facet includes all of the variability that could also be due to differences in Occasion. Therefore, differences in ratings due to an Occasion facet cannot be disentangled from the variance attributable to Rater Type. However, given the minimal amount of variance due to rater types, any variability due to differences in occasions is likely minor as well.

Second, the scores that the rater teams provide are, from a practical standpoint, the ratings that ultimately should be shown to be reliable. These scores represent the official ratings (after discrepancy adjudication) that programs use to gauge the relative strength of their assessment processes. Raters begin producing these official scores after the two day training. Therefore, at the time of the fully crossed G-studies, raters had

already read and rated fifteen programs in addition to the reports they used for calibration during training. Raters are very likely to be more consistent because of the extra practice. The results of this study reflect the consistency of raters who have had the experience of rating 15 programs and may therefore be slightly inflated when compared to the consistency of raters scoring their first few reports directly after training.

Finally, dependency issues among graduate student raters may have inflated the dependability estimates of the 2009-2010 fully-crossed G-study. During the first stage of the rating process, raters worked in teams to score a subset of programs, meeting at periodic intervals to examine their scores and adjudicate discrepancies. These meetings provided opportunities for raters to justify their scores and explain their rationale for scoring elements. Through this practice, raters had to come to an agreement within one point on the rubric scale, eliminating grossly discrepant scores. However, this adjudication process also provided an opportunity for pairs of raters to come to agreement on how to score various elements. Thus, during the second stage of the rating process (i.e., the fully-crossed design), graduate student rater scores may be dependent on the influence of their rater team partner during stage one.

This dependency issue only affects the graduate student ratings from the 2009-2010 study. These ratings were conducted by graduate students who had previously worked with one another in rater teams during stage one of the rating process. Because raters only worked with one other student during stage one, any dependency issues would be limited. During the 2010-2011 study, raters were all faculty members whereas the rater teams used in stage one of the rating process were comprised of one graduate

student and one faculty member. Therefore, dependency issues would not have impacted the consistency of scores during the 2010-2011 study involving faculty members.

It is important to recognize how design limitations may impact the interpretation of the results. For instance, given the structure of the current methodology and its inherent limitations, it may be more accurate to interpret the results of the fully-crossed designs as the dependability of any two "experienced" raters, with the term "experienced" referring to any rater who has completed training and rated approximately 15 assessment reports as part of a team. Although it is a nuanced change, the new interpretation is arguably more accurate and thus may provide stronger support for the resulting inferences.

To control for limitations, consider two alternate designs. First, the implementation of the fully-crossed design could occur before the official team ratings. In this first alternate design, by using both faculty and graduate student raters to estimate the consistency of ratings prior to separating everyone into teams, all three limitations are potentially diffused. For one, the occasion facet is eliminated by having graduate students and faculty members rate the same subset of programs concurrently. Second, the consistency of the ratings more accurately reflects how raters use the rubric from the outset of the rating process, instead of measuring the dependability after raters have scored 15 programs in teams. Third, rater dependencies are eliminated because raters have not had a chance to work with one another and adjudicate discrepancies.

Although the potential benefits of this first alternate design may sound appealing, there are drawbacks. Specifically, because raters would score a dozen reports independently, directly following training, it is very possible that they may develop their

own idiosyncratic opinions and rationales for scoring elements. When the individual raters are subsequently paired together and asked to rate programs as part of a team, it may be more difficult for them to come to consensus with their partners, especially on discrepant ratings. Therefore, although this first alternate design may provide an accurate estimate of dependability directly following training, it may have the adverse effect of lowering the consistency of the actual APT ratings.

As a second alternate design, the team ratings would occur prior to the fully-crossed design. However, to limit the potential biases that could develop among rater teams, raters would reconvene as a group after rating their first few program APTs. During this second calibration session, raters could ask questions and discuss issues that arose while rating their first few reports. Therefore, questions that occur among teams could be brought to the attention of the entire group of raters, alleviating the potential for team biases to occur.

Implications for the APT Validity Argument

Prior to this section of the discussion, the focus has been on interpreting the results relative to the four research questions posed in this study. However, the results presented here provide valuable evidence to strengthen the larger validity argument regarding APT scores. Beginning with the substantive stage of Benson's (1998) validation model, the discussion now returns to the initial assumptions that were made about the uses of the APT scores. The assumptions most impacted by the current study will be re-examined and those still in need of future attention will be discussed.

Additional evidence for the substantive stage. The assumptions in this stage were not addressed in the current research. Prior evidence effectively supports

assumptions one and two of the validity argument. Specifically, the rubric was created using the expertise of assessment professionals along with the consultation of appropriate literature to ensure that the breadth and scope of the assessment construct was adequately covered (Assumption One). Attention should be paid to new research surrounding the assessment construct as it evolves, but the initial work that was done to identify and construct the rubric should not be considered a priority of future validation arguments. Further, the rubric is an appropriate method for evaluating assessment (assumption two), which was supported by Fulcher et al.'s (2012) findings. As additional examples of meta-assessment become available, instrument developers may need to re-evaluate the number of behavioral anchors, anchor labels, and other aspects of the rubric, but it is evident that the majority of schools with available meta-assessments use a similar style and format to the APT rubric.

The third assumption, that the APT reports contain accurate information about the program's assessment report does not have formal supporting evidence. To gather evidence in support of this assumption, program assessments would need to be observed by trained raters who could then compare their observations to the contents of the report. Although it would take an extreme amount of resources to physically observe over 100 programs conduct assessment, it is possible to gather some evidence in support of this assumption. Program Assessment Support Services (PASS) consultants work with faculty members from various programs to help improve their assessment processes. These consultants work closely with several program faculty and often have first-hand knowledge of precisely what these programs do for their assessment. Thus, after the APT rating process has occurred, these PASS consultants could be used to compare the APT

report and subsequent ratings to what they observed through their interactions with the program. Although this strategy would not provide information about all programs' assessment processes, using consultants to confirm or refute the contents of the APT report may produce initial evidence to support the third validity assumption.

Additional evidence for the structural stage. The current research focused primarily on the assumptions of the structural stage in Benson's model. The results of the 2009-2010 study indicated that graduate students could rate assessment reports consistently relative to one another and with respect to the rubric's anchors. The results of the 2010-2011 study provide evidence that the ability of raters to produce relatively and absolutely dependable scores on the APT rubric can be generalized beyond graduate students to include faculty members. This evidence strongly supports structural stage assumptions four and five. Future research for these assumptions should focus on addressing the limitations from the current study to rule out possible conflicting causes for the results.

Additional evidence for the external stage. Formal evidence supporting the external stage of validation is still evolving. It is clear from Figure 1 that on average programs have improved in their element scores year over year. Fulcher et al. (2011) provided qualitative evidence to demonstrate that in many cases, the score improvements were due to real changes to the assessment process and not due to other causes. These findings support assumption six, which states that the rubric's anchors are sensitive to true differences in assessment quality. For future research, this assumption can also be supported by collecting known groups validity. That is, researchers can identify programs that have been credited with having strong assessment by external parties (e.g.,

professional accrediting agencies, Council for Higher Education Accreditation). If the APT rubric can differentiate programs with strong assessment from those with weaker assessments, the identified "strong" programs should receive high marks on the rubric.

To support assumption seven—the rubric scores are correlated with other appropriate measures of assessment—future studies can establish convergent validity by demonstrating that the APT rubric scores are appropriately correlated with other related constructs. For instance, research is currently underway to illustrate the relationship between APT change scores (i.e., the difference in APT scores from one year to the next) with the amount of contact a program makes with PASS consultants (Fulcher & Bashkov, in progress). To demonstrate convergent validity, programs with more contact with PASS staff (e.g., one-on-one meetings, workshops, help with data analysis) should have larger change scores year to year. However, because the measure of PASS "contact" may be challenging to define, future research should also focus on developing and identifying additional predictors of assessment, in order to determine whether the rubric can differentiate between various levels of program assessment.

The Impact of Meta-Assessment Research on the Assessment Field

The importance of meta-assessment in higher education assessment has been well-documented, but until now, little research has demonstrated validity evidence for a particular meta-assessment process. Although the current research has strengthened the validity argument of the APT meta-assessment process, one may question the broader policy implications of these findings. These implications exist at the institutional as well as the national level. For institutions, the meta-assessment process provides a process by which assessment results gain meaning. When programs have strong assessment

processes in place, their results provide better information to faculty and administrators about the quality of student learning. However, the methods by which assessment processes are judged must also be scrutinized and the uses of meta-assessment processes must be validated.

To help ensure that academic programs use sound processes to make inferences about student learning, assessment practitioners should question the methods by which assessment processes are evaluated. For instance, is the meta-assessment a checklist, a rubric, or some other form of assessment? What is the quality of feedback that can be ascertained through each type of measure? What are the criteria being used to evaluate assessment processes? Who is evaluating the quality of the assessment process and are they qualified to judge the veracity of assessment processes? Are the meta-assessment results used for summative or formative feedback? What types of decisions are being made from the results? These questions, and many others, form the foundation for identifying and validating the uses of the meta-assessment results.

Institutions must also consider the specific criteria practitioners use to evaluate assessment processes. Certain elements, such as the presence of clear and specific learning outcomes, are almost universally possible to evaluate across institutions. Other elements, such as ascertaining reliability estimates, may be more challenging for raters to judge or for programs to accomplish because of the resources available to the institution. Many schools may not have faculty with knowledge of various reliability estimates, nor do they have the resources to educate faculty about reliability. Further, practitioners may not have the expertise to provide meaningful feedback about the strength or quality of the assessment instrument's reliability. This is not to say that programs should ignore

reliability information, or seek to provide validity evidence in support of their instrument's uses. However, practicing good assessment can be a developmental process and not all institutions may have the resources, experience, or expertise to evaluate more advanced areas such as test score reliability. Thus, when creating a meta-assessment instrument, institutions must consider components of the assessment process that can be accomplished by programs just beginning assessment (e.g., outcomes, using results for program improvement), but that also fit within the capacity of the institution's resources.

To determine the essential components of the assessment process, an institution or program may rely on the standards set by professional and regional accrediting agencies. For example, the Southern Association of Colleges and Schools – Commission on Colleges (SACSCOC) requires schools to "[1] identify expected outcomes, [2] assess the extent to which [they] achieve these outcomes, and [3] provide evidence of improvement based on analysis of results" (SACSCOC, 2012, p. 27). Accrediting standards often reflect the minimal level needed to demonstrate effectiveness. Thus, an institution with limited resources may rely on these three elements outlined by SACSCOC for a logical foundation to its meta-assessment criteria.

Other institutions with cultures of assessment and ample resources may set higher standards for their academic programs. James Madison University, for example, has one of the largest assessment centers in the United States with nine full-time faculty, over a dozen graduate students, and several administrative staff, all of whom provide assessment support to the broader campus community. Given these resources, the elements of the APT rubric reflect criteria that academic programs should strive to achieve. Whereas some schools may only be able to evaluate three to five elements that cover broad aspects

of the assessment process, other schools such as JMU have the personnel to effectively evaluate a larger number of criteria. Thus, these meta-assessments can focus on more minute details of the assessment process where expert feedback can greatly enhance the overall quality of assessment.

Ultimately, the criteria chosen for an institution's meta-assessment should reflect elements that align with the school's culture of assessment, professional best practices, and resources available to the institution. Meta-assessments that contain criteria that cannot be met by faculty and cannot be reliably rated by evaluators are more likely to lead to untrustworthy results and scores that are not useful to assessment practitioners and administrators. In cases where a foundation of strong assessment practice has not been implemented, or resources for assessment support are limited, it may become necessary to sacrifice a wide array of challenging criteria for a select few that lead to useful, consistent feedback.

A third policy implication for institutions is that in order for meta-assessment scores to facilitate improved assessment processes, the results must be reliable. Strong reliability estimates of meta-assessments indicate the relative clarity of the instrument, quality of the evaluators, and trustworthiness of the scores. Institutional administrators need to consider whether faculty, staff, or students (or a combination of all three) are most qualified to evaluate assessment reports. Further, administrators need to consider the number of raters to use in order to rate programs consistently. This decision impacts the amount of time and resources that will be devoted to the rating process. If too few raters are used, the scores may be undependable and lose meaning. If too many raters are used, valuable resources are wasted. Reliable ratings indicate that evaluators can come to

similar conclusions about the relative strengths and weaknesses of a program's assessment. Without precise ratings, the usefulness of the scores is weakened. If administrators wish to use results to examine real gains in assessment quality over time, imprecise ratings may inflate negligible gains or mask real growth across programs. Faculty may use feedback resulting from imprecise ratings in ways that actually damage their assessment process more than it helps. In short, poor reliability calls into question the quality of the instrument and the expertise of the raters, both of which are valuable keys to forming a case for validity.

National Policy Implications

The implications for validating meta-assessment processes do not reside solely at the institutional level. In fact, there are several national policy implications for conducting meta-assessment. Regional accrediting bodies such as SACSCOC conduct reviews of colleges and universities and evaluate the quality of the institutions relative to the agencies' standards. Although schools are evaluated on the same standards by a team of higher education professionals, each institution is reviewed by a different team. Specifically, these accrediting agencies attempt to evaluate the degree to which schools assess the learning outcomes of educational programs. Thus, within the accreditation process, evaluators are conducting institution-wide meta-assessments. The question then becomes, what evidence exists to support the decisions accreditation evaluators make regarding the quality of institutions' assessment? What criteria do evaluators use to determine compliance? How reliable are these decisions?

The argument for evidence supporting the meta-assessment processes by accrediting agencies is not simply philosophical. The decisions made by the evaluator

team to deem institutions in compliance with not just the assessment standards but all accreditation standards have high-stakes consequences; institutions not in compliance with any standards can potentially lose their federal funding. Therefore, it is imperative that reviewers interpret all accreditation standards consistently across institutions so that schools are rewarded and reprimanded based upon real differences in quality, rather than the subjective tendencies of their specific evaluators. In short, accrediting agencies have a responsibility to show stakeholders evidence that the reaffirmation process leads to intentional, reliable, and valid decisions that support the intended purpose of accreditation; meta-assessment provides a pathway for these agencies to demonstrate that validity evidence.

Regional accreditors are not the only professional organizations with a need to validate their meta-assessment processes. The New Leadership Alliance for Student Learning and Accountability (NLA) is an "advocacy-focused organization" committed to improving student learning in American undergraduate education by supporting voluntary efforts by higher education institutions to conduct meaningful assessment (NLA, 2012a). As one of its initiatives, the NLA provides the Excellent Practice in Student Learning Assessment institutional certification program (NLA, 2012b). This program acknowledges institutions that practice a high standard of quality regarding assessment. As part of becoming certified in this program, institutions submit a report of their assessment processes. Using rubrics, a team of evaluators from across the country assesses the degree to which these programs meet the established criteria; those institutions that become certified are then recognized for their high assessment standards. This approach to meta-assessment has the potential to be generalizable across the

spectrum of American higher education as NLA leaders attempt to set forth standards for assessment that transcend state and regional policies. Even though institutional participation is voluntary, as the NLA agenda gains traction with federal and state legislatures, institutions who fail to meet these standards may face negative consequences from lawmakers, making the need for trustworthy results from the NLA a necessity.

Given the potential for high-stakes decisions to come about as a result of the NLA's certification process, institutional leaders are likely to question the validity of the claims the NLA makes regarding its meta-assessment scores. An institution that receives poor scores on the NLA's meta-assessment faces damage to its reputation in the eyes of its stakeholders and it could become the target of state scrutiny. Because the NLA hopes to generalize these findings to a national audience, the organization must demonstrate that the process is appropriate for a national audience. One of the best ways to illustrate the wide-reaching application of this certification process is by demonstrating the reliability of the ratings provided to participating institutions. The team that evaluates institutional assessment performance is comprised of a diverse group of higher education professionals from across the country. To demonstrate to policymakers and other stakeholders that certification in the NLA's program is a national achievement of sound assessment practice, members of the evaluation team must come to similar conclusions about the quality of an institution's assessment practice. Without such consistency, what worth do the ratings actually hold? What does it mean to be certified if the experts tasked with evaluating the quality of an institution's commitment to assessment cannot agree on what quality assessment looks like? To address public scrutiny, NLA leaders must be

able to answer these and other questions regarding the validity of their meta-assessment process.

Efforts from regional accreditors and groups like the NLA to hold institutions to higher standards regarding assessment illustrate the ever-increasing role of meta-assessment in conversations about the assessment movement. Demands for accountability in higher education by state and federal policymakers are not likely to subside any time soon. Therefore, it is imperative that institutions develop methods of evaluating the quality of their assessment practices and, as importantly, gather evidence supporting their meta-assessment processes. The APT meta-assessment validity argument can provide a framework for assessment practitioners and stakeholders in higher education seeking to evaluate the quality of assessment processes at the academic program, institutional, and national level.

Appendix A

Assessment Progress Template

For Annual Academic Degree Program Reporting

Introduction and Purpose:
The purpose of this template is to provide the most current assessment-related information for each of JMU's academic programs. A separate template will be completed for each academic major program offered at JMU. With this information, James Madison University will have information to share with both internal and external constituents about the quality of all academic programs.

How to fill out the APT:
Objectives - Please provide your academic program's learning goals and objectives. Describe the process by which the objectives receive faculty review. Which, if any, of your objectives were modified, deleted, or added in the last year?

Course/Learning Experiences - Provide the linkage between your program's goals and objectives and their instructional delivery via your curriculum. This can be demonstrated with a matrix that lists the goals and objectives by the courses that address each.

Evaluation/Assessment Methods - Provide a listing of the systematic methods and procedures for gathering information about achievement of your goals and objectives. Additionally, specify the ***expected*** student achievement results. Please also describe the process for systematic data collection. Finally, describe the measurement properties of the assessment method, such as reliability and validity.

Objective Accomplishments/Results - Provide a description of your program's assessment results for the last two years. Provide an interpretation of the program's assessment results. What do these results mean for you and your faculty? In your interpretation, refer back to your objectives/instructional methods and expectations of results.

Dissemination- Describe how your assessment results are shared with your faculty and others concerned with your program. Illustrate how your assessment results are incorporated in the planning and governance structure of your program.

Uses of Evaluation/Assessment Results and Actions Taken. Demonstrate how the program's assessment results have been used to contribute to program improvement and enhanced student learning and growth. Examples of program actions taken might include modification and/or additions to learning objectives, curriculum revisions, instructional delivery changes, changes in course sequencing, or increased emphasis on specific skill development. Additionally, explain any changes to the assessment process you have made this year or plan to make in the coming year.

Appendix B

Assessment Progress Template Rubric

*Assessment Progress Template (APT) Evaluation Rubric, Version 3.0**

I. Student-centered learning objectives

Beginning 1	Developing 2	Good 3	Exemplary 4	Score
A. Clarity and Specificity				
No objectives stated.	Objectives present, but with imprecise verbs (e.g., know, understand), vague description of content/skill/or attitudinal domain, and non-specificity of whom should be assessed (e.g., "students")	Objectives generally contain precise verbs, rich description of the content/skill/or attitudinal domain, and specification of whom should be assessed (e.g., "graduating seniors in the Biology B.A. program")	All objectives stated with clarity and specificity including precise verbs, rich description of the content/skill/or attitudinal domain, and specification of whom should be assessed (e.g., "graduating seniors in the Biology B.A. program")	
B. Orientation				
No objectives stated in student-centered terms.	Some objectives stated in student-centered terms.	Most objectives stated in student-centered terms.	All objectives stated in student-centered terms (i.e., what a student should know, think, or do).	

II. Course/learning experiences that are mapped to objectives

Beginning 1	Developing 2	Good 3	Exemplary 4	Score
No activities/courses listed.	Activities/courses listed but link to objectives is absent.	Most objectives have classes and/or activities linked to them.	All objectives have classes and/or activities linked to them.	

**Note. Only ratings labels – 1(Beginning), 2(Developing), 3(Good) – have been modified from Version 2.0; In all other respects Version 3.0 is identical to 2.0*

III. Systematic method for evaluating progress on objectives

Beginning 1	Developing 2	Good 3	Exemplary 4	Score
A. Relationship between measures and objectives				
Seemingly no relationship between objectives and measures.	At a superficial level, it appears the content assessed by the measures matches the objectives, but no explanation is provided.	General detail about how objectives relate to measures is provided. For example, the faculty wrote items to match the objectives, or the instrument was selected "because its general description appeared to match our objectives."	Detail is provided regarding objective-to-measure match. Specific items on the test are linked to objectives. The match is affirmed by faculty subject experts (e.g., through a backwards translation).	
B. Types of Measures				
No measures indicated	Most objectives assessed primarily via indirect (e.g., surveys) measures.	Most objectives assessed primarily via direct measures.	All objectives assessed using at least one direct measure (e.g., tests, essays).	
C. Specification of desired results for objectives				
No a priori desired results for objectives	Statement of desired result (e.g., student growth, comparison to previous year's data, comparison to faculty standards, performance vs. a criterion), but no specificity (e.g., students will grow; students will perform better than last year)	Desired result specified. (e.g., our students will gain ½ standard deviation from junior to senior year; our students will score above a faculty-determined standard). "Gathering baseline data" is acceptable for this rating.	Desired result specified and justified (e.g., Last year the typical student scored 20 points on measure x. The current cohort underwent more extensive coursework in the area, so we hope that the average student scores 22 points or better.)	

D. Data collection & Research design integrity				
No information is provided about data collection process or data not collected.	Limited information is provided about data collection such as who and how many took the assessment, but not enough to judge the veracity of the process (e.g., thirty-five seniors took the test).	Enough information is provided to understand the data collection process, such as a description of the sample, testing protocol, testing conditions, and student motivation. Nevertheless, several methodological flaws are evident such as unrepresentative sampling, inappropriate testing conditions, one rater for ratings, or mismatch with specification of desired results.	The data collection process is clearly explained and is appropriate to the specification of desired results (e.g., representative sampling, adequate motivation, two or more trained raters for performance assessment, pre-post design to measure gain, cutoff defended for performance vs. a criterion)	

E. Additional validity evidence				
No additional psychometric properties provided.	Reliability estimates (e.g., internal consistency, test-retest, inter-rater) provided for most scores, although reliability tends to be poor (<.60). Or, author states how efforts have been made to improve reliability (e.g., raters were trained on rubric).	Reliability estimates provided for most scores, most scores are marginal or better (>.60).	Reliability estimates provided, most scores are marginal or better (>.60). Plus, other evidence given such as relationship of scores to other variables and how such relationship strengthens or weakens argument for validity of test scores.	

IV. Results of program assessment

Beginning 1	Developing 2	Good 3	Exemplary 4	Score
A. Presentation of results				
No results presented	Results are present, but it is unclear how they relate to the objectives or the desired results for the objectives.	Results are present, and they directly relate to the objectives and the desired results for objectives but presentation is sloppy or difficult to follow. Statistical analysis may or may not be present.	Results are present, and they directly relate to objectives and the desired results for objectives, are clearly presented, and were derived by appropriate statistical analyses.	

B. History of results				
No results presented	Only current year's results provided.	Past iteration(s) of results (e.g., last year's) provided for some assessments in addition to current year's.	Past iteration(s) of results (e.g., last year's) provided for majority of assessments in addition to current year's.	
C. Interpretation of Results				
No interpretation attempted	Interpretation attempted, but the interpretation does not refer back to the objectives or desired results of objectives. Or, the interpretations are clearly not supported by the methodology and/or results.	Interpretations of results seem to be reasonable inferences given the objectives, desired results of objectives, and methodology.	Interpretations of results seem to be reasonable given the objectives, desired results of objectives, and methodology. Plus, multiple faculty interpreted results (not just one person). And, interpretation includes how classes/ activities might have affected results.	

V. Documents how results are shared with faculty/stakeholders

Beginning 1	**Developing 2**	**Good 3**	**Exemplary 4**	**Score**
No evidence of communication	Information provided to limited number of faculty or communication process unclear.	Information provided to all faculty, mode and details of communication clear.	Information provided to all faculty, mode and details of communication clear. In addition, information shared with others such as advisory committees, other stakeholders, or to conference attendees	

VI. Documents the use of results for improvement

Beginning 1	**Developing 2**	**Good 3**	**Exemplary 4**	**Score**
A. **Improvement of programs regarding student learning and development**				
No mention of any improvements.	Examples of improvements documented but the link between them and the assessment findings is not clear.	Examples of improvements (or plans to improve) documented and directly related to findings of assessment. However, the improvements lack specificity.	Examples of improvements (or plans to improve) documented and directly related to findings of assessment. These improvements are very specific (e.g., approximate dates of implementation and where in curriculum they will occur.)	

B. **Improvement of assessment process.****				
No mention of how this iteration of assessment is improved from past administrations.	Some critical evaluation of past and current assessment, including acknowledgement of flaws, but no evidence of improving upon past assessment or making plans to improve assessment in future iterations.	Critical evaluation of past and current assessment, including acknowledgement of flaws; Plus evidence of some moderate revision, or general plans for improvement of assessment process.	Critical evaluation of past and current assessment, including acknowledgement of flaws; both present improvements and intended improvements are provided; for both, specific details are given. Either present improvements or intended improvements must encompass a major revision.	

**Note, if the assessment has received predominantly Exemplary ratings, then that program will automatically receive a "3" for VI B.

References

Alexander, L., Clinton, B., & Kean, T. H. (1986). *Time for results: The governors' 1991 report on education.* Washington, D.C.: The National Governors' Association.

American Association for Higher Education. (1993). *Principles of good practice for assessing student learning.* Washington, DC: Author.

American Evaluation Association. (2004). *Guiding principles for Evaluators (*2nd Ed.). Fairhaven, MA: Author. Retrieved from: http://www.eval.org/Publications/aea06.GPBrochure.pdf (Original work published 1994)

Association of American Colleges & Universities. (2002). *Criteria for recognizing "good practice" in assessing liberal education.* Washington: Author. Retrieved from: http://www.aacu.org/gex/paa/assessment.cfm

Association of American Colleges & Universities. (2008). *Our students' best work: A framework for accountability worthy of our mission (*2nd Ed.). Washington, DC: Author.

Australian Universities Teaching Committee. (2002). *Core principles of effective assessment.* Melbourne, Australia: Author. Available online: http://www.cshe.unimelb.edu.au/assessinglearning/05/index.html

Banta, T. W. (2002). Characteristics of effective outcomes assessment: Foundations and examples. In Banta, T. W. (Ed.), *Building a scholarship of assessment* (pp. 261-283). San Francisco, CA: Jossey-Bass.

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

Boyer, C.M.. & Ewell, P. T. (1988). *State-based approaches to assessment in undergraduate education: A glossary and selected references*. Denver, CO: Education Commission of the States.

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice*, 11, 27-34.

Brennan, R. L. (2001). *Manual for urGENOVA.* Iowa City, IA: Iowa Testing Programs, University of Iowa.

Bresciani, M. J. (2003). Expert-driven assessment: Making it meaningful. *Educause Center for Applied Research (ECAR) Research Bulletin*, 2003(21), 1-13.

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2009). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs.* Sterling, VA: Stylus.

California State University, Long Beach Institute for Teaching and Learning. (1993). *Academic Challenges: Student Outcomes Assessment.* Washington, D.C.: Fund for the Improvement of Post-Secondary Education.

Cooksy, L. J. (1999). The Meta-evaluand: The evaluation of project TEAMS. *American Journal of Evaluation, 20(1)*, 123-136.

Cottarelli, C. & Escolano, J. (2004). *Assessing the assessment: A critical look at the June 2003 assessment of the United Kingdom's five tests for euro entry*. International Monetary Fund Working Paper.

Council of Regional Accrediting Commissions. (2004). *Regional accreditation and student learning: A guide for institutions and evaluators.* Retrieved from: http://www.anokaramsey.edu/resources/pdf/assessment/assessmentguidecrac.pdf

Crick, G. E., & Brennan, R. L. (2001). *GENOVA: A generalized analysis of variance system* (Fortran IV computer program and manual.) Dorchester, MA: University of Massachusetts at Boston, Computer Facilities.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, Theory, and Public Policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* (52), 281-302.

Davis, B. G. (1989). Demystifying assessment: Learning from the field of evaluation. *New Directions for Higher Education* 67, 5-20.

Driscoll, A., & De Noriega, D. C. (2006). *Taking Ownership of Accreditation: Assessment Processes That Promote Institutional Improvement and Faculty Engagement*. Sterling, VA: Stylus Publishing.

Erwin, T. D. (1991). *Assessing Student Learning and Development: A Practical Guide for College Faculty and Administrators.* San Francisco: Jossey-Bass Publishers.

Ewell, P. T. (1988). Implementing assessment: Some organizational issues. In T. W. Banta (Ed.), *Implementing Outcomes Assessment: Promise and Perils* (New Directions for Institutional Research No. 59, pp. 15-28). San Francisco: Jossey-Bass Publishers.

Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T.W. Banta & Associates (Eds.), *Building a scholarship of assessment,* (pp. 3-25). San Francisco: Jossey Bass Publishers.

Fitzpatrick, J. L. (2004). Exemplars as case studies: Reflections on the links between

theory, practice, and context. *American Journal of Evaluation, 25(4),* 541-559.

Fong Bloom, M. (2010, September). Peer review of program assessment efforts: One strategy, multiple gains. *Assessment Update,* 22(5), 5-7, 16.

Fulcher, K. H. & Bashkov, B. M. (2012). *The accountability of an assessment office.* Unpublished manuscript, Department of Graduate Psychology, James Madison University, Harrisonburg, VA.

Fulcher, K. H., Coleman, C., Dabback, W., Haraway, D., Orem, C. D., & Rodgers, M. R. (18 September 2011). *Measuring and promoting assessment improvement*. Invited presentation for the Center for Assessment and Research Studies, Harrisonburg, VA.

Fulcher, K. H. & Orem, C. D. (2010). Evolving from quantity to quality: A new yardstick for assessment. *Research and Practice in Assessment*, *4*(1), 1-10. Retrieved from http://www.virginiaassessment.org/rpa/5/FulcherandOrem.pdf

Fulcher, K. H., Sundre, D. L, & Russell, J. A. (2009). The assessment progress template rubric. *Product developed by the Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA.*

Fulcher, K. H., Swain, M. S., & Orem, C. D. (2012 January/February). Expectations for assessment reports: A descriptive analysis. *Assessment Update,* 24(1), 1-2, 14-16.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Grasso, P. G. (1999). Meta-evaluation of an evaluation of reader focused writing for the veterans benefits administration. *American Journal of Evaluation,* 20(2), 355-370.

Greater Expectations Project on Accreditation and Assessment. (2004). *Taking*

responsibility for the quality of the baccalaureate degree. Washington, DC: National Academy Press.

Hatfield, S. (2009). Assessing your program-level assessment plan. Idea Paper #45. *The IDEA Center*. Retrieved from www.theideacenter.org

Huba, M.E. and Freed, J.E. (2000). Applying principles of good practice in learner-centered assessment. In M. E. Huba & J. E. Freed (Eds.), *Learner-Centered Assessment on College Campuses: Shifting the focus from teaching to learning* (pp. 65-90). Needham Heights, MA: Allyn and Bacon.

Jacobi, M., Astin, A., & Ayala, F. (1987). *College student outcomes assessment: A talent development perspective (ASHE-ERIC Higher Education Report No. 7*). Washington, DC: Association for the Study of Higher Education.

Johnson, R., Prus, J., Andersen, C. J., & El-Khawas, E. (1991). Assessing assessment: An in-depth status report on the higher education assessment movement in 1990. *Higher Education Panel Report No. 79*. Washington, DC: American Council on Education.

Joint Committee on Standards for Educational Evaluation (1981). *Standards for Evaluation of Educational Programs, Projects, and Materials.* New York: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd Ed.). Thousand Oaks, CA: Sage Publications, Inc.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards* (3rd Ed.). Thousand Oaks, CA: Sage Publications, Inc.

Kane, M. T. (1992) An argument-based approach to validity. *Psychological Bulletin*, 112,

527-535.

McDonald, B. (2010). Improving learning through meta-assessment. *Active Learning in Higher Education,* 11(2). 119-129.

Marchese, T. J. (December 1987). Third down, ten years to go. *AAHE Bulletin* 38, 10-13.

Messick, S. J. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher,* 18, 5-11.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author. Retrieved from: http://www.natd.org/Code_of_Professional_Responsibilities.html

National Governors' Association. (2007). *Higher education accountability for student learning*. Retrieved from: http://www.nga.org/files/live/sites/NGA/files/pdf/0702HIGHERED.PDF

New Leadership Alliance for Student Learning and Accountability (2012a). *Mission*. Retrieved from: http://www.newleadershipalliance.org/who_we_are/about_us/

New Leadership Alliance for Student Learning and Accountability (2012b). *Institutional Certification.* Retrieved from: http://www.newleadershipalliance.org/what_we_do/excellent_practice_in_student_learning_assessment/

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

Orem, C. D. & Fulcher, K. H. (2011). *Demonstrating the utility of program-level meta-assessments: An application of generalizability theory.* Manuscript submitted for publication.

Ory, J.C. (1992). Meta-Assessment: Evaluating assessment activities. *Research in Higher*

Education, 33(4), 467-481.

Palomba, C. A., & Banta, T. W. (1999). *Assessment Essentials: Planning, Implementing and Improving Assessment in Higher Education.* San Francisco, CA: Jossey-Bass Publishers.

Peters, T.A. (2005). Current opportunities for the effective meta-assessment of online reference services. *Library Trends,* 49(2), 334-349.

Peterson, M. W. & Einarson, M. K. (2001). What are colleges doing about student assessment? *Journal of Higher Education* 72(6), 629-669.

Reindl, T. & Reyna, R. (July 2011). *From information to action: Revamping higher education accountability systems.* National Governors' Association Center for Best Practices. Retrieved from: http://20.132.48.254/PDFS/ED522081.pdf

Rossman, J. E., & El-Khawas, E. (1987). *Thinking about assessment: Perspectives for presidents and chief academic officers*. Washington, DC: American Council on Education and the American Association for Higher Education.

Scott-Little, C., Hamann, M., & Jurs, S. G. (2002). Evaluations of after-school programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation* 23(4), 387-419.

Shavelson, R . & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

Shavelson, R. (2010). *Measuring college learning responsibly: Accountability in a new era.* Stanford, CA: Stanford University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education,* 19, 405-450.

Smits, P. A. & Champagne, F. (2008). An assessment of the theoretical underpinnings of practical participatory evaluation. *American Journal of Evaluation* 29(4), 427-442.

Southern Association of Colleges and Schools – Commission on Colleges. (2012). *The principles of accreditation: Foundations for quality enhancement* (5th Ed.). Decatur, GA: Author.

Steen, L. A. (1999). Assessing assessment. In B. Gold et al (Eds.), *Assessment practices in undergraduate mathematics* (pp. 1-6). Washington: Mathematical Association of America.

Study Group on the Conditions of Excellence in American Higher Education. (1984). *Involvement in learning: Realizing the potential of American higher education* (ED Publication No. 246 833). Retrieved from: http://www.eric.ed.gov/PDFS/ED246833.pdf

Stufflebeam, D.L. (1968). *Evaluation as enlightenment for decision-making*. Columbus, Ohio: Evaluation Center, Ohio State University.

Stufflebeam, D. L. (2000). The methodology of metaevaluation as reflected by metaevaluations by the Western Michigan University Evaluation Center. *Journal of Personnel Evaluation in Education* 14(1), 95-125.

Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation* 22(2), 183-210.

Suskie, L. A. (2000). Fair assessment practices: Giving students equitable opportunities to demonstrate learning. *AAHE Bulletin* 52, 7-9.

Suskie, L. A. (2006). *What is "good" assessment? A variety of perspectives.* Retrieved from: http://outcomes.lbcc.edu/pdf/GdAssessPerspectives.pdf

Suskie, L. A. (2009). *Assessing student learning: A common sense guide (*2nd ed.). San Francisco: Jossey-Bass.

Tyler, R. W. (1950). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

VanLeeuwen, D. M. (1997). Assessing reliability of measurements with generalizability theory: An application to inter-rater reliability. *Journal of Agricultural Education*, 38(3), 36-42.

Walker, D.A. (1999). A model for assessing assessment activities. *College Student Journal, 33*(3), 439-443.