2012

A Study of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunities

Timothy Chow
Indiana State University

VITA

Timothy Kin Chuen Chow

EDUCATION

Degree	Field	Institution	Year
PhD	Technology Management (Quality Systems)	Indiana State University	2012
MBA	Business Administration	Indiana State University	1992
BS	Quality and Decision Sciences	Indiana State University	1993
BS	Business Administration / Management Information Systems	Indiana State University	1991

EMPLOYMENT HISTORY

Rose-Hulman Institute of Technology – Director of Institutional Research 1999-Present

- Responsible for providing support to institutional management, operations, planning, policy-formation and decision-making through various tasks and research activities.
- Lead data gathering and survey research, quantitative and qualitative analysis, internal and external reporting, and benchmarking process.

Rose-Hulman Institute of Technology – Assistant Dean for Institutional Research 1997-1999
Rose-Hulman Institute of Technology – Information Resource Manager 1996-1997
Indiana State University – Institutional Research Coordinator 1995-1996
Indiana State University – Testing Associate 1994-1995

PROFESSIONAL AFFILIATIONS

Association for Institutional Research (AIR)
Indiana Association for Institutional Research (INAIR)
Oman Academic Accreditation Authority (OAAA) External Reviewer
Overseas Chinese Association for Institutional Research (OCAIR)

PUBLICATION

Rogers, G.M., & Chow, T. (2000). Electronic portfolios and the assessment of student learning. *Assessment Update*, *12*(1), 4-6.

A STUDY OF A COMPUTER-AIDED PERFORMANCE RATING PROCESS AND THE ASSOCIATED PROCESS IMPROVEMENT OPPORTUNITIES

A Dissertation

Presented to

The College of Graduate and Professional Studies

College of Technology

Indiana State University

Terre Haute, Indiana

In Partial Fulfillment

of the Requirements for the Degree

PhD in Technology Management

---

by

Timothy Chow

August 2012

Keywords: academic quality, inter-rater and intra-rater reliability, performance rating process, quality standards, technology management

COMMITTEE MEMBERS

Committee Chair: M. Affan Badar, Ph.D.

Chair and Associate Professor of Applied Engineering and Technology Management

Indiana State University

Committee Member: George R. Maughan, Ed.D.

Professor and Director of the PhD in Technology Management Program

Indiana State University

Committee Member: John W. Sinn, Ed.D.

Chair and Professor of Engineering Technologies

Bowling Green State University

Committee Member: Ronald C. Woolsey, Ph.D.

Program Coordinator and Professor of Industrial Management

University of Central Missouri

ABSTRACT

Academic quality measurement through assessing student learning outcomes targets the crux of teaching and learning activities undertaken by higher education institutions. With the proliferation of information and instructional technology, authentic assessments of student academic performance by utilizing a computer-aided performance rating process offer promises of more precise and actionable information to educators for making informed quality improvement decisions on curricular changes and a viable alternative to standardized tests. This pilot research study examined the validity and reliability of a computer-aided performance rating process. Furthermore, this research offered information to the educational community on the feasibility of adapting a scalable performance measurement solution and its implications for improving academic quality.

ACKNOWLEDGMENTS

I am grateful to the following individuals and organizations for their encouragement and support for completing this research:

• Dissertation Committee members: Dr. M. Affan Badar, Dr. George R. Maughan, Dr. John W. Sinn, and Dr. Ronald C. Woolsey

• Drs. Ming Zhou and Todd Waggoner, and other Consortium PhD Program faculty and staff members

• The College of Graduate and Professional Studies at Indiana State University

• My immediate supervisor Dr. Julia M. Williams and former supervisor Dr. Gloria M. Rogers, and colleagues at Rose-Hulman Institute of Technology

• Professor Emeriti of Mechanical Engineering Dr. C. Mallory North at Rose-Hulman Institute of Technology and Mrs. Sonya North

• My wife Rita, my children Charissa and Christopher, my parents Mr. and Mrs. John and Ada Chow, and my brothers Gabriel and Godfrey Chow

• My friends Aaron G. and Monna M. Brown, and Zach Yoder

• All my colleagues in the field of institutional research and all my friends I have known for all these years

My greatest thanks are to my Lord and Savior Jesus Christ for giving me strength and endurance to overcome many challenges and complete this dissertation!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Defining and measuring educational or academic quality remain challenges facing higher education institutions today. Measuring academic quality is a multidimensional challenge that comes with a wide array of educational outcomes and expectations; therefore, it is not likely that any single measure would be sufficient to account for the diversity of factors and constituents involved in the educational process (Saunders, 2007; Vaughn, 2002). Aside from lacking quality standards accepted and implemented by higher education institutions across the nation, what defines and measures the outcomes of quality education is currently still hotly debated and contested in the public arena (Dill and Soo, 2005; Blackmur, 2008; Rollins, 2011). Similar phenomena can be seen around the world as in Europe with the remaining challenges of implementing a framework of common degree qualifications through the Bologna Accord or Bologna Process (CHEPS, 2010), in Australia and other parts of the world with maintaining quality outcomes and academic standards (Shah and Brown, 2009).

With heightened public scrutiny concerning access to and funding of higher education and on the accountability of educational outcomes (Arum and Roksa, 2011; Hacker and Dreifus, 2010; Newton, 2000; Rothchild, 2011), colleges and universities affiliated with various higher education organizations, such as institutional and program accrediting bodies and professional societies, are pressured to make progress of student learning outcomes public (Prados, Peterson,

& Lattuca, 2005; Shavelson, 2007; Sullivan & Thomas, 2007). In addition to the traditional input measures--incoming students' prior achievement, resources and spending, and output measures--retention, graduation, placement and loan default rates among others, the current accountability movement has imposed further requirements of higher education institutions to demonstrate knowledge, skills gain and even changes of attitudes of their students along the educational process (Higher Learning Commission, 2003; State Higher Education Executive Officers, 2005; U.S. Department of Education, 2006; Ewell, 2008; Thomson & Douglass, 2009; Williams, 2010; Council for Higher Education Accreditation, 2011). Ongoing efforts are exerted to draw common ground among higher education institutions for articulating expected educational outcomes and mastery level required of college degree holders. These efforts and quality standard development activities can be seen more prominently among academic programs in professional and technical fields, such as accounting, architecture, dentistry, engineering, law, medicine, pharmacy, psychology, teaching, and veterinary (Carpenter, et al., 2008).

For general or common educational outcomes, the Lumina Foundation (2011) has recently released a proposed version of a Degree Qualifications Profile to serve as a framework for defining the expected educational outcomes of associate, bachelor's and master's degree holders, regardless of their majors or fields of study. Rather than relying on typical credit counting for describing degree expectations at the institutional level, the Degree Qualifications Profile proposes common learning outcomes for the above-mentioned three levels of degrees. The proposed outcomes are organized in five broad areas of competencies as illustrated in Figure 1: specialized knowledge, broad or integrated knowledge, applied learning, intellectual skills, and civil learning (Lumina Foundation, 2011). The intent is to offer reference points for students

and broader audience on acquiring field-specific knowledge and competencies at the respective degree levels.

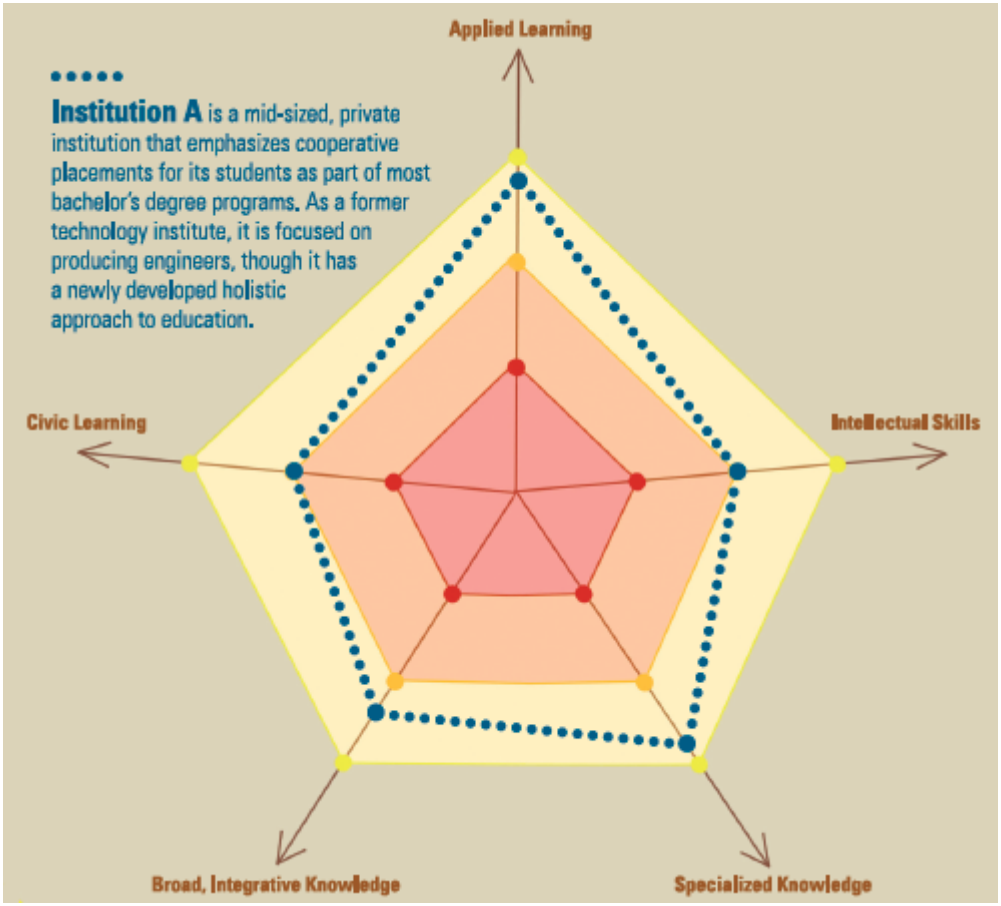


Figure 1. Five areas of competencies of the Degree Qualifications Profile. Reprinted from *The Degree Qualifications Profile* (p. 7). Copyright 2011 by the Lumina Foundation for Education, Inc. Reprinted with Permission.

For program specific educational outcomes such as civil engineering, the American Society of Civil Engineers released its second edition of the Civil Engineering Body of Knowledge (BOK) for the 21st Century in 2008. This publication offers current definitions of the educational outcomes, that is, knowledge, skills, and attitudes expected of college graduates entering the practice of civil engineering (American Society of Civil Engineers, 2008). In the

publication, the 24 educational outcomes are divided into three categories: foundational, technical, and professional; and they are presented in the form of Bloom's Taxonomy with varying levels of achievement being specified for each outcome. Crosswalk tables of outcomes among the first edition of the BOK, the second edition of the BOK and the Engineering Accreditation Commission (EAC) of the ABET, Inc. are included in Table 1 to illustrate the general relationships among these outcomes (American Society of Civil Engineers, 2008). Furthermore, the outcome rubric for the BOK is provided in Appendix D to indicate the required levels of achievement or competence for each outcome and the roles of education and prelicensure experience (American Society of Civil Engineers, 2008).

With more concrete steps taken and progress made toward consensus building in defining both common and program specific learning outcomes, the natural next step is to identify viable options that best measure achievement of the stated outcomes and yield relevant information for assessing and improving academic quality. Aside from using grades to assess student learning and surveys to capture self-reported experiences and satisfaction in courses and programs, one of the popular measurement choices of general education outcomes is the utilization of standardized tests on reading, critical thinking and problem solving skills (Shavelson, 2007). Likewise, professional and technical degree programs are utilizing standardized subject field tests or licensure examinations to assess students' competencies on specific topics relevant to the disciplines.

Table 1 ASCE Civil Engineering Body of Knowledge for the 21st Century.

ABET Outcomes[a]	*BOK1 Outcomes*[b]	*BOK2 Outcomes*[c]
(a) Mathematics, science, engineering	1. Technical core	1. Mathematics 2. Natural sciences 5. Materials science 6. Mechanics
(b) Experiments	2. Experiments	7. Experiments
(c) Design	3. Design	9. Design 10. Sustainability
	3. Design	12. Risk/uncertainty
(d) Multidisciplinary teams	4. Multidisciplinary teams	21. Teamwork
(e) Engineering problems	5. Engineering problems	8. Problem recognition and solving
(f) Professional and ethical responsibility	6. Professional and ethical responsibility	24. Professional and ethical responsibility
(g) Communication	7. Communication	16. Communication
(h) Impact of engineering	8. Impact of engineering	11. Contemporary issues and historical perspectives
(i) Lifelong learning	9. Lifelong learning	23. Lifelong learning
(j) Contemporary issues	10. Contemporary issues	11. Contemporary issues and historical perspectives 19. Globalization
(k) Engineering tools	11. Engineering tools	8. Problem recognition and solving
	12. Specialized area related to civil engineering	15. Technical specialization
Program Criteria for Civil and Similarly Named Engineering Programs	13. Project management, construction, and asset management	13. Project management
	14. Business and public policy	17. Public policy 18. Business and public administration
Program Criteria for Civil and Similarly Named Engineering Programs	15. Leadership	20. Leadership 22. Attitudes
EAC/ABET Criterion 5[d]	EAC/ABET Criterion 5[d]	3. Humanities 4. Social sciences
Program Criteria for Civil and Similarly Named Engineering Programs	Program Criteria for Civil and Similarly Named Engineering Programs	14. Breadth in civil engineering areas

a) Short names[12]

b) Short names of outcomes appearing in the BOK1 report,[3] pp. 24–29

c) Short names from this report, Table 1, page 16

d) General education component

[a] General relationships are presented, not one-to-one mapping.

Although standardized tests could provide comparable information across higher education institutions on topics with established norms or commonly agreed educational standards being measured through these tests, the benchmarking information yielded often is not helpful in providing information with precision for improving and refining current curricular offerings and student learning experience. As a result, a movement of using a more holistic, criterion-referenced, rubric-based performance review or rating approach has gained traction in recent years in response to the need for obtaining information for improving academic quality and refining the educational process (Davies & Le Mahieu, 2003; Association of American Colleges and Universities, 2005). In fact, similar performance rating process has been in place for over decades in classroom assessment of student performance and it comes in various forms, such as capstone performances, oral examinations, product and performance evaluations, and portfolios (Palomba and Banta, 1999). Palomba and Banta (1999) defined the performance rating process (as an approach opposed to standardized tests or surveys) is intended to evaluate students' knowledge, skills, and development in an authentic manner. The primary difference in the current adaptation of the performance rating approach is at the broader level, so that common quality standards can be applied across and beyond class sections while a systematic measurement process can be used to judge the quality of student performance in a more objective fashion (Walvoord, 2004). What it all boils down to is that the performance rating process offers an authentic and actionable way for higher education institutions to satisfy accountability demands on quality assurance of educational programs, as well as to improve their core functions of promoting teaching and learning excellence. With the proliferation of information and instructional technology, drawbacks of the typical or traditional performance rating process such as storage and information retrieval could be remediated to expedite the steps involved in the

measurement process and allow for wider adaptation of the process by higher education institutions.

With general acceptance of the performance rating process in classroom settings (Brown, 2004; McMartin, McKenna & Youssefi, 2000), this research aimed at examining the validity and reliability of a computer-aided or non-traditional performance rating process and exploring the feasibility of expanding such a process for measuring program and institutional student learning outcomes and improving academic quality. The demonstrated student performances for this research were gathered through selected courses using embedded assignments. These courses offered students' opportunities to develop and demonstrate competencies associated with the specific institutional student learning outcomes of Rose-Hulman Institute of Technology. The RH3 Communication outcome and the RH4 Cultural and Global Awareness outcome that are mapped respectively to the Criteria 3g and 3h of the ABET's General Criteria for Baccalaureate Level Programs, as shown in Table 2, were chosen for the research.

Table 2 ABET Criterion 3g and 3h outcomes mapped to Rose-Hulman RH3 Communication and RH4 Cultural and Global Awareness Institutional Student Learning Outcomes

ABET Criterion 3 Selected Program Outcomes	Rose-Hulman Institute of Technology Selected Institutional Student Learning Outcomes	
	RH3 Communication	RH4 Cultural/Global Awareness
g. Communications	√	
h. Global Society		√

Aside from being recognized as part of the engineering program requirements by ABET, Inc., these two student outcomes are also reflected in many general education outcome statements published by higher education institutions across the nation. This research offered an illustration of using a computer-aided performance rating process to document these outcomes

and satisfy part of the ABET accreditation requirements related to student outcomes. The reliability of the computer-aided process was analyzed through the derived inter-rater and intra-rater reliability indices. The performance rating process diagrams are illustrated in Appendix E and the formulae for deriving these rater agreement indices are included in Appendix F.

Statement of the Problem

The problem for this study is to assess the validity and reliability of a computer-aided performance rating process, which may serve as a viable option and offer relevant information for measuring and improving educational or academic quality.

As demands for evidence of student learning from higher education institutions continue to rise, so do the demands for a scalable solution that would meet both the accountability requirements and quality improvement needs. As common expectations of student learning outcomes are emerging, the performance rating process is gaining more attention due to its design for authentic assessment of student competencies. However, it is unclear if the drawbacks of the traditional performance rating process can be minimized by utilizing available information and instructional technology, so that broader adaptation of such a process would become feasible to higher education institutions.

Significance of the Study

The research goals were to identify the strengths and weaknesses associated with the computer-aided performance rating process and to improve the overall quality of such a process. The findings offered information to the community that examines student outcomes about the feasibility to adapt a scalable performance rating process for obtaining actionable information in making curricular improvement changes. A promising scalable and objective performance rating process would enable all constituents to engage in meaningful dialogues of academic quality and

provide the means to measure academic quality in a purposeful way beyond quality assurance. Without such a solution, higher education institutions would likely be left with limited options such as standardized tests to serve as the primary device that could address only part of the academic quality question.

Research Objectives

The problem for the study is addressed through three objectives:

1. Assess the validity of the computer-aided performance rating process.
2. Examine the reliability of the computer-aided performance rating process.
3. Investigate opportunities to expand the computer-aided performance rating process into a scalable solution for enhancing the efficiency of the educational outcome measurement process.

Assumptions and Limitations

The following assumptions were made for this study:

1. The validity and reliability of the generally accepted traditional performance rating process will be assumed to facilitate the assessment of validity and reliability of the computer-aided process through comparative analysis.
2. The performance criteria of the selected student learning outcomes for the research are assumed to possess face validity; that is, the stated criteria for the outcomes are the appropriate specifications of what they are supposed to measure.
3. The rubrics or evaluation standards for judging or rating demonstrated competencies through work samples are assumed to possess content validity; that is, the relevant content knowledge is well represented in the stated criteria for judging performance.

The following limitations were inherent to this study:

1. This study is experimental in nature and is limited to a sample of raters involved in measuring academic performance of students on selected learning outcomes.

2. The rating results used for rater agreement and other analyses in this pilot study are based on relatively small samples of student work collected and assessed at a single institution.

3. There is an ongoing effort in higher education to define core competencies and common expectations of both general and program specific educational outcomes. However, the results of this effort are not yet widely available for adaptation among higher education institutions.

4. This study is focusing on summative measures of academic quality; that is, direct assessments of student learning on college graduates' knowledge, skills, and abilities.

Definition of Terms

The following are definitions of terms that are being referenced in this study:

ABET

Formerly known as the Accreditation Board for Engineering and Technology, Inc. and the formal name is now ABET, Inc. since 2005. ABET is a nonprofit, non-governmental organization that accredits college and university programs in the disciplines of applied science, computing, engineering, and engineering technology (ABET, n.d.a).

ASCE

The American Society of Civil Engineers. ASCE represents more than 140,000 members of the civil engineering profession worldwide and is America's oldest national engineering society (American Society of Civil Engineers, n.d.)

ASCE Civil Engineering Body of Knowledge (BOK)

This publication offers current definitions of the educational outcomes, that is, knowledge, skills, and attitudes expected of college graduates entering the practice of civil engineering (American Society of Civil Engineers, 2008).

Authentic assessment

Performance assessments call upon the examinee to demonstrate specific skills and competencies, that is, to apply the skills and knowledge they have mastered (Stiggins, 1987).

Bloom's Taxonomy

It is a classification of educational objectives into three domains: cognitive, affective, and psychomotor and varying knowledge and skill levels ranging from lowest order or foundational to highest order levels (Bloom, 1956).

Bologna Process

Formerly known as the Bologna Accords. The Bologna Process initially was an initiative for establishing the European Higher Education Area (EHEA) to ensure more comparable, compatible and coherent systems of higher education in Europe. Since then, the Bologna Process moves towards a new phase focusing on a reduction of the implementation discrepancies in the countries forming the EHEA (Leuven/Louvain-la-Neuve Communiqué, 2009).

Chi-square test for homogeneity

The Chi-square test statistic is equal to the squared difference between the observed and expected frequencies, divided by the expected frequency in each cell of the table, summed over all cells of the table. The test statistic approximately follows a chi-square distribution with 1 degree of freedom. For evaluating differences in portions between two groups, the Z test of two

proportions is equivalent to the Chi-square test for the difference between two proportions (Berenson, Levine & Krehbiel, 2012).

Content validity

It is based on the extent to which a measurement reflects the specific intended domain of content (Carmines & Zeller, 1991).

Degree Qualifications Profile

It is a framework for defining and measuring the general knowledge and skills that individual students need to acquire in order to earn degrees at various levels, such as associate, bachelor's and master's degrees (Lumina Foundation, 2011).

Effective

An effective process produces output that conforms to customer requirements. The lack of process effectiveness is measured by the degree to which the process output does not conform to customer requirements, that is, by the level of defect of the output (Lewis & Smith, 1994).

Effectiveness

The state of having produced a decided or desired effect; the state of achieving customer satisfaction (Lewis & Smith, 1994).

Efficiency

A measure of performance that compares output production with cost or resource utilization as in number of units per employee per hour or per dollar (Lewis & Smith, 1994).

Efficient

An efficient process produces the required output at the lowest possible (minimum) cost. That is, the process avoids waste or loss of resources in producing the required output. Process efficiency is measured by the ratio of required output to the cost of producing that output. This

cost is expressed in units of applied resource, such as dollars, hours, energy, etc. (Lewis & Smith, 1994).

Engineering Accreditation Commission (EAC)

The EAC leads and conducts its accreditation activities related to the engineering discipline. The EAC is responsible for reviewing educational programs and rendering decisions about engineering accreditation (ABET, n.d.).

Face validity

It is concerned with how a measure or procedure appears. Does it seem like a reasonable way to gain the information the researchers are attempting to obtain? Does it seem well designed? Does it seem as though it will work reliably? Unlike content validity, face validity does not depend on established theories for support (Fink, 1995).

High trait prevalence

High trait prevalence refers to a condition when the prevalence of a given response is very high.

Inter-rater reliability

For the purpose of this dissertation, inter-rater reliability refers to the consensus estimate between two raters (Stemler, 2001).

Intra-rater reliability

It is a metric for rater's self-consistency in the scoring of subjects. (Gwet, 2008b).

Juried review

It refers to one form of direct assessment of performance by using a jury or a panel of experts.

Liberal Education and American's Promise (LEAP)

LEAP is a national initiative of the Association of American Colleges and Universities (AAC&U) to promote a set of essential learning outcomes fostered through liberal education (Association of American Colleges and Universities, 2005).

Norming activity

For the purpose of this dissertation, a norming activity refers to the establishment of a common threshold in applying the scoring rubrics between two or among three or more raters for judging performance.

Outcome

For the purpose of this dissertation, the focus is on student outcome. Student outcomes describe what students are expected to know and able to do by the time of graduation. These relate to the knowledge, skills, and behaviors that students acquire as they progress through the program (ABET, n.d.b).

Performance criteria

Specific, measurable statements identifying the performance(s) required to meet the outcome; confirmable through evidence (ABET, n.d.b).

Performance Rating

It is the output of an assessment process to indicate the level of achievement according to the given performance scale.

Primary Trait Analysis (PTA)

PTA is an assessment method for establishing explicit criteria for grading in a classroom setting and allowing for performance assessment at program and/or institutional level (Walvoord, 2004).

Raters

For the purpose of this dissertation, raters are being referred to content experts who assess demonstrated performances and assign performance ratings.

Rater agreement analyses

Studies involved in applying statistical methods to analyze agreement among performance ratings from two or more raters. They can be used to evaluate a new rating system or instrument, validate a new rating system or instrument, aid in decisions about combining performance ratings of two or more raters to obtain evaluations of suitable accuracy (Uebersax, 2000).

Rating results

For the purpose of this dissertation, the rating results refer to the pass or fail ratings assigned by raters during the performance rating process.

Reliability

The degree to which a measure yields consistent results (Mueller, 2011).

RosE-Portfolio System (REPS)

A computer-aided performance rating system that is developed at Rose-Hulman Institute of Technology in 1996 as a method for assessing student learning outcomes.

RosEvaluation Tool (RET)

A new version of the computer-aided performance rating tool that interfaces with course or learning management system to facilitate online assessment of student learning outcomes.

Rubric

A description about the expected level(s) of performance for instructional or grading/scoring purposes.

Standardized tests

Objective assessments such as short answer, completion, multiple-choice, true-false, and matching tests are structured tasks that limit responses to brief words or phrases, numbers or symbols, or selection of a single answer choice among a given number of alternatives (Linn & Miller, 2009).

Summative measures

The gathering of information at the conclusion of a course, program, or undergraduate career to improve learning or to meet accountability demands (Leskes, 2002).

Tuning USA

It is methodology based on the Bologna Process aiming to enhance the transparency regarding what a degree represents in the US by exploring and defining curricula learning outcomes of selected subject areas (The Institute for Evidence-Based Change, 2010).

University and College Accountability Network (U-CAN)

It is a consumer-information initiative developed by the National Association of Independent Colleges and Universities (NAICU) to give students and parents online access to information on nonprofit, private colleges and universities in a common format (University & College Accountability Network (U-CAN), n.d.).

Valid Assessment of Learning in Undergraduate Education (VALUE)

This project is a part of the AAC&U's Liberal Education and America's Promise (LEAP) initiative to develop institutional level rubrics for the selected essential learning outcomes (Association of American Colleges and Universities, n.d.b).

Validity

The degree to which a certain inference from a measure is appropriate and meaningful (Mueller, 2011).

Voluntary System of Accountability (VSA)

It is a voluntary initiative developed by the public 4-year universities and sponsored by the Association of Public and Land-grant Universities (APLU) and the Association of State Colleges and Universities (AASCU) to supply clear, accessible, and comparable information on the undergraduate student experience to important constituencies through a common web report – the College Portrait (Voluntary System of Accountability Program, n.d.).

CHAPTER 2

REVIEW OF LITERATURE

The review of literature will focus on service quality, its relationship to the current state of academic quality measurement, and will provide general descriptions of the performance rating process. The first section reviews the definitions and fundamental concepts under quality and service quality. The following section examines how these definitions and concepts are applicable and related to the measurement of academic quality. The third section describes relevant issues surrounding quality standards development and the performance rating process. A brief summary is provided as a conclusion for this chapter.

Quality and Service Quality

The word quality can be defined in many different ways. Some people might think of quality as degree of excellence, while others might consider quality as free of defects in products or services, or the features or price associated with them. Evans and Dean (2003) offered the following responses to the definition of quality by managers of 86 firms in the eastern United States: perfection, consistency, eliminating waste, speed of delivery, compliance with policies and procedures, providing a good and usable product, doing it right the first time, delighting or pleasing customers, and total customer service and satisfaction.

Furthermore, Evans and Dean (2003) suggested that “fitness for use” or “meeting or exceeding customer expectations” seems to be one of the most popular definitions among the

survey responses. Likewise, as cited by Evans and Dean (2003), the American National Standards Institute (ANSI) and the American Society for Quality (ASQ) define quality as "the totality of features and characteristics of a product or service that bears on its ability to satisfy given needs." The customer-focus definition of quality resonates well with the ones offered by the quality gurus, such as Juran and Deming.

Quality System

A quality system can be defined as a collection of policies, standards, processes, and resources that are documented, implemented and maintained to provide a framework for examining practices and achieving quality objectives. Bucher (2004) offers a succinct account of the basic premise of a good quality system: Say what you do, do what you say, record what you did, check the results, and act on the difference, which is essentially an implementation of the Shewhart's or Deming's Plan-Do-Check-Act Cycle.

Service Quality

Lewis and Booms (1983) defined service quality as a measure of how well a delivered service matches the customer's expectations. From the quality planning perspective, Zeithaml, Parasuraman and Berry (1990) suggested that quality of a service could be examined through five dimensions: tangibles, reliability, responsiveness, assurance, and empathy. From the quality control and improvement perspectives, popular quality tools such as quality function deployment (QFD) and statistical process control (SPC) can be applied to assess, monitor, and improve service quality (Maguad, 2009; Tsung, Li, & Jin, 2008).

In addressing quality service issues, the key challenge is not knowing exactly what the customer's needs and service performance standards are, since these expectations are to be defined by the customers who could have difference preferences. Furthermore, most of the

measurement approaches are indirect measures, such as customer satisfaction surveys. Unlike product quality that can be quantified with direct measures, customer needs and service performance standards can be difficult to measure.

Service Quality and Academic Quality

In ways similar to defining the terms of quality and service quality, the term "academic quality" could mean different things to different audiences. Unlike product quality where tangible products manufactured are typically being assessed against standardized requirements and design specifications, higher education institutions are service organizations involved in educational processes and deliver intangible outputs of which quality is being measured by both the customer's subjective expectations and their diverse needs and past experiences. Since student learning is at the core of the missions and purposes of higher education institutions, special attention should be paid to direct measures of learning and its role in defining educational or academic quality. The RAND Corporation (Chun, 2002) identified four primary measures in determining higher education quality in his research: actuarial data, ratings of institutional quality, student surveys, and direct measures of student learning.

Actuarial Data

The actuarial data include mostly input and output measures of the educational process, such as retention and graduation rates, composition of student body, credentials of faculty members, levels of external funding, and admission test scores, etc. The underlying assumption in utilizing actuarial data for judging academic quality is based on the assumption that better inputs are more likely to yield better outputs in the end of the educational process.

Ratings of Institutional Quality

Institutional quality ratings include surveys of faculty members and college administrators on their perceptions and opinions about academic quality and reputation. Rankings of higher education institutions, such as the Academic Ranking of World Universities published by the Center for World-Class Universities and the Institute of Higher Education of Shanghai Jiao Tong University-China, the Times Higher Education World University Rankings published by TSL Education Limited, and the America's Best Colleges published by U.S. News and World Report, are typical examples of measures of perceived academic quality through actuarial data analysis and institutional quality ratings among peers in the field.

Student Surveys

Self-reported information obtained through surveys and/or interviews on students' collegiate experiences, satisfaction and self-assessment of educational gains, such as the National Survey of Student Engagement (NSSE), is another popular approach, attempting to link educational quality with student learning. With NSSE, the research suggests that student engagement represents two critical features of academic quality. The first is the amount of time and effort students devote to their studies and other educationally purposeful activities. The second is how the institution deploys its resources and organizes the curriculum and other learning opportunities to get students to participate in activities that research studies have linked to enhancing student learning.

Direct Assessments of Student Learning

Course grades analysis, comprehensive examinations, standardized tests, performance tasks, open-ended tests, evaluations of student projects and portfolios of student work are examples of direct measures of student learning. Although the direct assessments of student

learning is a logical way to assess academic quality, Chun (2002) identified barriers such as cost and lack of consensus of what should be measured and how best to assess student learning.

Academic Quality and Quality Standards Development

With the current landscape of higher education facing significant challenges that mimic similar situations faced by American manufacturing companies of the late 1970s and early 1980s, the strategic importance of quality is once again being recognized by higher education institutions as a powerful tool for retaining competitive strength and ensuring survival in light of massive educational budget cuts under the volatile economic climate. At the same time, additional pressures on higher education institutions come from increased regional, national and global competitions and heightened expectations from constituent groups for increased productivity (Hunt Jr. and Tierney, 2006). For assuring quality in higher education, accreditation agencies of educational institutions and programs are demanding quality systems be maintained by higher education institutions, so that performances can be assessed according to the standards, policies, resources and procedures specified in the respective quality systems (Weldy and Turnipseed, 2010).

Here in America, quality standards based on the prominent Malcolm Baldrige National Quality Award (MBNQA) Criteria or ISO 9000 Standards, such as the Academic Quality Improvement Program (AQIP) Categories of the Higher Learning Commission as shown in Table 3, provide frameworks for both implementing and assessing quality within higher education institutions (College of DuPage, n.d.).

Table 3 Comparing the AQIP and Baldrige Criteria by College of DuPage

AQIP Criteria	Baldrige Criteria
Helping Students Learn Accomplishing Other Distinctive Objectives Supporting Institutional Operations Building Collaborative Relationships	Educational and Support Process Management • Education Support Processes • Partnering Processes
Understanding Students and Other Stakeholders Needs	Student and Stakeholder Focus • Student and stakeholder satisfaction and relationships
Valuing People	Faculty and Staff Focus • Faculty and staff education, training, and development • Faculty and staff well-being and satisfaction
Leading and Communicating	Leadership • Public responsibility and citizenship
Measuring Effectiveness	Information and Analysis • Analysis of organizational performance
Planning Continuous Improvement	Strategic Planning • Strategic deployment

However, consensus is still lacking in defining common standards and agreeing on approaches for assessing the outcomes of quality education. Criticisms of current accreditation systems are appearing in news articles and media headlines: pushing for a minimum level of quality, focusing narrowly on process as with the ISO 9000:1994, and ignoring significant outcomes or results of quality education in terms of student learning (Wergin, 2005; Gillen, Bennett, & Vedder, 2010). Simply put, current quality measures and performance indicators of

academic quality either do not address the perennial question of what students learn at higher education institutions or do not satisfy diverse constituents' expectations: employers on employee readiness, legislatures regarding accountability of public funds, professional societies and higher education institutions interested in sustainability and future development, and the general public concerning investment and choice for educational services (Kuo, 2006).

In responding to these criticisms, ongoing efforts are made by higher education organizations, accreditation agencies and professional societies to find common ground in articulating expected program educational outcomes and mastery level of knowledge and skills required of college graduates. Two evolving and collaborative efforts in developing qualifications frameworks for defining commonly expected college and discipline level learning outcomes are led by the Lumina Foundation (2011) called the Degree Qualifications Profile and Tuning USA. Aside from the Degree Qualifications Profile initiative described in the previous chapter, the Tuning USA (Figure 2) is based on similar methodology of the Bologna Process, aiming to enhance the transparency regarding what a degree represents by exploring and defining curricula learning outcomes of selected subject areas (Adelman 2009; The Institute for Evidence-Based Change 2010).

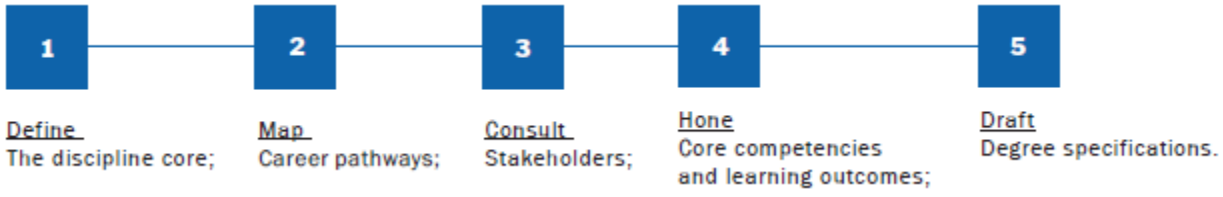
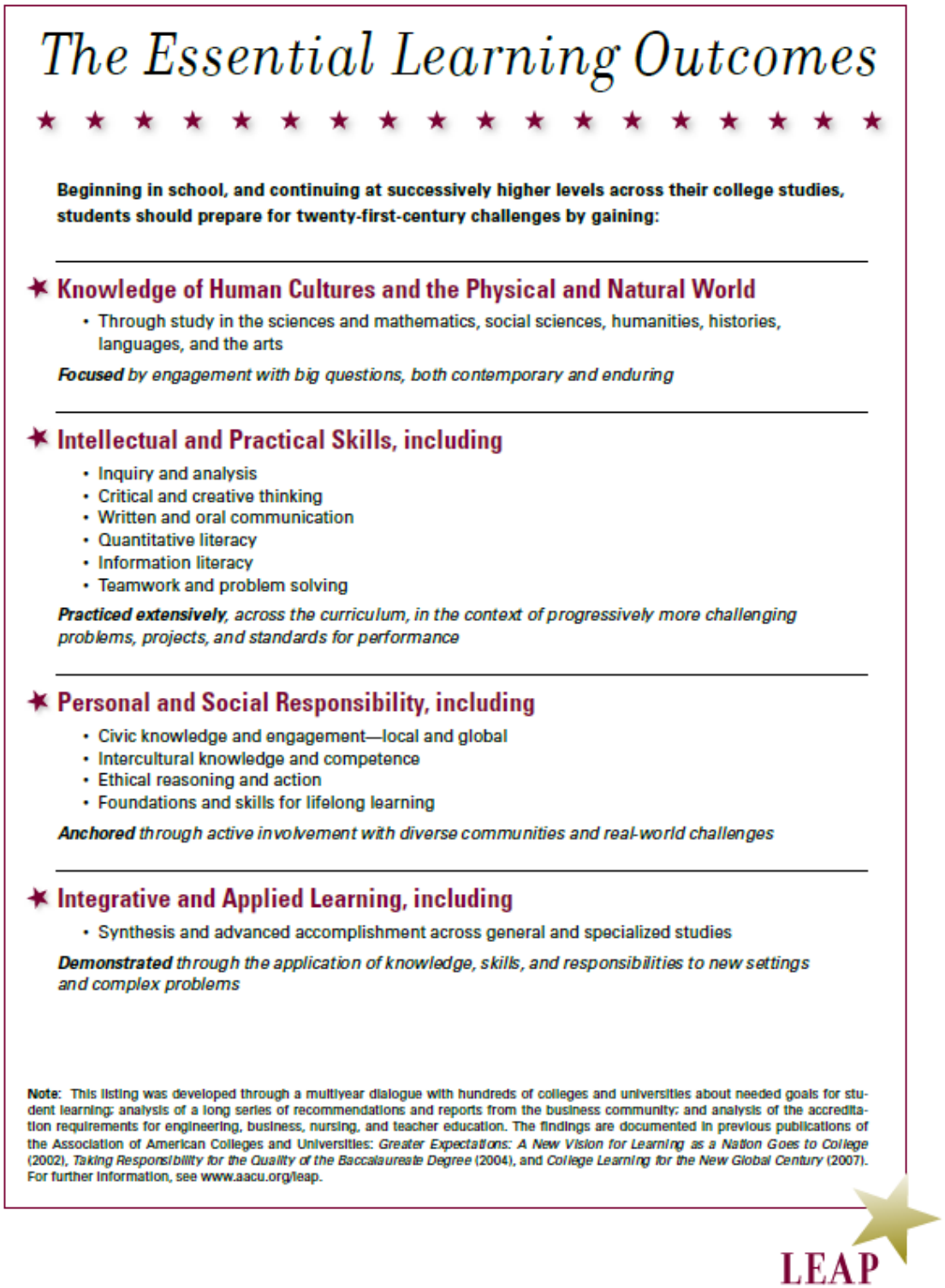


Figure 2. Tuning USA Five Steps Process

Five steps involved in the "Tuning" process in establishing degree specifications in the Tuning Educational Structures guide by IEBC 2010. Adapted from "Tuning Educational Structures" by the Institute for Evidence-Based Change (IEBC), 2010, p.3.

Moreover, many public and private higher education institutions are participating in the respective voluntary accountability systems, such as the Voluntary System of Accountability (VSA) and the University and College Accountability Network (U-CAN), to disclose online comparable and qualitative campus information and performance measures to the prospective students and their parents, and the public. For measuring student learning and reporting gains in broad cognitive skills at institutional level, VSA participants are required to provide student experience survey results from one of the four chosen national surveys: the College Student Experiences Questionnaire, the College Senior Survey, the National Survey of Student Engagement, or the University of California Undergraduate Student Experience Survey. Aside from survey results, VSA participants are required to publish the passing rates of licensure examinations and other national program-specific direct assessment results via the online College Portraits. Furthermore, the VSA participants are strongly encouraged to provide either their institutional student learning outcomes assessment results or test scores from one of the following three selected national standardized tests: the Collegiate Assessment of Academic Proficiency, the Collegiate Learning Assessment, or the Measure of Academic Proficiency and Progress. On the other hand, U-CAN participants face fewer prescriptive requirements, and they may voluntarily disclose student learning outcomes measures on their online profiles.

For contributing to the national dialogue on assessment of student learning, the Association of American Colleges and Universities (AAC&U) has launched an initiative called the Liberal Education and America's Promise (LEAP). The LEAP initiative is intended to promote a set of national college learning outcomes, as shown in Figure 3, and accompanying rubrics established through a project called the Valid Assessment of Learning in Undergraduate Education (VALUE).



The Essential Learning Outcomes

Beginning in school, and continuing at successively higher levels across their college studies, students should prepare for twenty-first-century challenges by gaining:

Knowledge of Human Cultures and the Physical and Natural World

- Through study in the sciences and mathematics, social sciences, humanities, histories, languages, and the arts

Focused *by engagement with big questions, both contemporary and enduring*

Intellectual and Practical Skills, including

- Inquiry and analysis
- Critical and creative thinking
- Written and oral communication
- Quantitative literacy
- Information literacy
- Teamwork and problem solving

Practiced extensively*, across the curriculum, in the context of progressively more challenging problems, projects, and standards for performance*

Personal and Social Responsibility, including

- Civic knowledge and engagement—local and global
- Intercultural knowledge and competence
- Ethical reasoning and action
- Foundations and skills for lifelong learning

Anchored *through active involvement with diverse communities and real-world challenges*

Integrative and Applied Learning, including

- Synthesis and advanced accomplishment across general and specialized studies

Demonstrated *through the application of knowledge, skills, and responsibilities to new settings and complex problems*

Note: This listing was developed through a multiyear dialogue with hundreds of colleges and universities about needed goals for student learning: analysis of a long series of recommendations and reports from the business community; and analysis of the accreditation requirements for engineering, business, nursing, and teacher education. The findings are documented in previous publications of the Association of American Colleges and Universities: *Greater Expectations: A New Vision for Learning as a Nation Goes to College* (2002), *Taking Responsibility for the Quality of the Baccalaureate Degree* (2004), and *College Learning for the New Global Century* (2007). For further information, see www.aacu.org/leap.

LEAP

Figure 3. A proposed set of "Essential Learning Outcomes"

Reprinted from Essential Learning Outcomes, n.d., Retrieved June 9, 2010, from http://www.aacu.org/leap/documents/EssentialOutcomes_Chart.pdf.

The set of proposed essential learning outcomes is targeting the typical general education outcomes currently being examined via standardized tests, such as inquiry and analysis, critical thinking, written communication, and quantitative literacy. In addition, student learning outcomes, such as teamwork and problem solving, information literacy, civic knowledge and engagement, intercultural knowledge and competence, ethical reasoning and action, foundations and skills for lifelong learning, and synthesis and advanced accomplishment across general and specialized studies are included in the proposed comprehensive set of essential outcomes. The AAC&U has emphasized that the VALUE project builds on a philosophy of learning assessment that privileges authentic assessment of student work and shared understanding of student learning outcomes on campuses over reliance on standardized tests administered to samples of students outside of their required courses.

A set of rubrics for the identified fifteen essential student learning outcomes is currently being field tested and refined, while the implementations of authentic performance measures using the VALUE rubrics such as electronic portfolios are being examined (Association of American Colleges and Universities, 2005). Furthermore, the AAC&U has issued statements on ongoing collaboration with the Lumina Foundation to set expected standards for the meaning of the degree through the LEAP and the Degree Qualifications Profile initiatives (Association of American Colleges and Universities, 2011). These efforts to develop quality standards are representative of ongoing attempts at the national level by higher education organizations to respond to the needs of diverse constituents for information on academic quality.

In Europe, similar efforts to develop quality standards in higher education area began with the signing of the voluntary Bologna Declaration in 1999 with 30 initial participating

countries to 47 countries now forming the European Higher Education Area (CHEPS, 2010). The Bologna Accords, now being referred to as the Bologna Process, aims to:

- Strengthen the competitiveness and attractiveness of the European higher education,
- Foster student mobility and employability through the introduction of common qualifications frameworks, with an emphasis on learning outcomes, and
- Promote European cooperation in quality assurance.

With noticeable progress made in establishing a common degree level system for undergraduates (bachelor's degree) and graduates (master's and doctoral degrees) among participating countries, significant effort is still required of a large number of participating countries to implement the common qualifications framework at both institutional and national levels beyond the anticipated 2010 completion target. Furthermore, similar to the American higher education system, there are remaining challenges for European higher education institutions and the European Higher Education Area to articulate program-learning outcomes and identify performance measures of the intended learning outcomes (CHEPS, 2010).

Academic Quality and Performance Rating Process

When considering measurement options, standardized tests remain one of the most common approaches that higher education institutions are using to assess students' reading, critical thinking and problem solving skills and to compare samples of their students' performances against selected peer institutions or national norms (Shavelson, 2007). With the known shortfalls of standardized tests such as artificial time limit for problem-solving, financial cost, representation, test taking strategies of participants, and at times, the relevance of topics to what is being taught in the classroom; the emerging choice of measuring educational quality through direct assessments of student learning has begun to gain attention in the educational

community (Buu, 2003; Shulman, 2007). This alternative criterion-referenced, rubric-based academic performance rating approach as described in the previous chapter is not new. Outside of academia, the performance rating process is commonly applied for evaluating aircrew and pilot performance, judging job applicants' performance, as well as for assessing gymnasts' performance (Alsmadi, 2005; Deaton, et al, 2007). In academia, Bresciani (2006) traced the "inquiry-based notion of 'how well do we know what we are doing is working'" to the juried reviews dated back to 1063 CE at the University of Bologna, and its presence at the Universidad de Salamanca, Spain, in 1230.

As an alternative to standardized tests, performance rating process is based on one of the popular classroom assessment choices for judging academic performance of students in individual courses. This variant of the classroom assessment approach is to make program and even institutional student learning outcomes and attributes explicit to students and other evaluators, so common quality standards and measurement process may be applied to judge the quality of student learning in a more systematic and objective manner. Buu (2003) addressed the emerging trend in performance rating process regarding its potential for measuring higher-order thinking better than multiple-choice questions and the importance of rater effect measurement, as well as offered an overview of statistical measures, such as Cohen's Kappa, for evaluating performance rating process. On the other hand, Palomba and Banta (1999) identified some possible limitations to the performance rating process, such as time and labor costs, availability of empirical research on validity and generalizability of results, etc. Some of these challenges could be overcome with utilization of information and instructional technology and with clearly defined outcomes and scoring rubrics (Lombardi 2008). A study conducted by Mazor et al. (2007) further indicate the importance of having clear and shared expectation of performance

and rubrics (what and how to judge) when applying the performance rating process. Furthermore, Roch et al. (2009) recommended attention should be given to the role of rater training, such as Frame of Reference, and the number of performance dimensions to be rated as conditions for increasing rater agreement.

Walvoord (2004) offers a glimpse into the inner workings of a criterion-referenced, rubric-based performance rating process called Primary Trait Analysis (PTA) that resembles a more structured fashion of the typical grading process for judging student performance in a classroom setting and beyond. The essential steps in PTA involve instructors identifying key performance criteria or traits to be learned and demonstrated by students according to course or learning objectives, specific performances or assignments for demonstrating desired competencies, levels of achievement for each primary trait for judging the quality of student performances, and providing anchors or descriptions of the performance expectations at each level of achievement. A rubric in this performance measurement approach refers to the information piece that lists the primary traits and the grading standards associated with each level of achievement. The rubric helps articulate both the learning outcomes and expected performances that sharpen students' focus and provide feedback for guiding their steps in achieving the learning objectives. In addition, the rubric offers a more consistent and objective way for instructors to judge student performances and adds transparency to the performance measurement process that incorporates appropriate embedded course assessment results for meeting broader assessment needs across class sections and even for informing program and institutional assessment efforts. Schamber and Mahoney (2006) offered examples to illustrate how rubrics can be utilized for the authentic assessment of group critical thinking in a first-year,

core general education course, as well as to demonstrate the use of assessment data for revising a curriculum, improving instruction, and enhancing student learning.

Rogers and Chow (2000) shared their experience in applying instructional and information technology to leverage the advantages of performance rating process with an electronic portfolio system called the RosE-Portfolio System (REPS). REPS is designed for judging the quality of student outcomes that aims at satisfying the requirements of both the institutional and program accreditation agencies. At the same time, information collected through the system is utilized for assessing program effectiveness and informing curricular changes. Among a variety of student learning assessment methods such as course grades, questionnaires and surveys, standardized tests, and other qualitative methods, Rose-Hulman's faculty members selected portfolios for the following reasons: richness of quality information about students in a broad range of outcome areas that could be obtained, enhanced validity with direct assessment approach in measuring student performance that reflects educational offerings, sensitivity to time commitment of both students and faculty members, and the engaging nature of active learning involvement by students throughout the performance measurement process. Rogers and Williams (2001) highlighted one of the distinct advantages, with a systems approach to measure performance: the adaptability of such process for use in individual classes as well as for use at programs or institutional levels by utilizing a common language of assessment and rubrics. When designing the process with special attention to identifying critical components and requirements such as the focus and scope of assessment, measurable learning objectives, roles of students/faculty/other constituents, and specific measurement steps and feedback, Rogers and Williams reported that the requirements analysis step has promoted the efficiency and validity of the measurement process. Furthermore, they emphasized the goal for incorporating

information and instructional technology is to enhance the performance rating process by making it more efficient and user-friendly. Efficiency is achieved by minimizing efforts associated with the measurement process such as access, store, view, and rate student performances and overall information management on learning outcomes. Moreover, students, faculty members and other constituents can gain asynchronous access to REPS online and at their convenience throughout the process.

Traditional Performance Rating Process

The traditional performance rating process involves a team of two or more raters. After completing a "norming" exercise, the raters evaluate all identical work samples for a particular outcome independently and throughout the entire rating process (Goldie et al., 2004). The norming exercise involves a group review of the pre-defined rubrics and individually assigned ratings on the selected anchor documents, that is, representative student work samples, early on in the rating process to establish the threshold for assigning ratings for the remainder of the process. The purpose for the norming exercise is to help raters gain better understanding of the scoring rubrics and reach consensus on how to apply the rubrics consistently throughout the process. Tamanini (2008) conveyed the importance of the norming exercise or Frame-of-Reference (FOR) training in particular and rater training in general through his background research that these activities have been shown to increase rating accuracy. Ottolini and her colleagues (2007) offered similar conclusion regarding the importance of rater training for reducing the variability of assigned ratings associated with the performance rating process. Typically, the average ratings or weighted scores among raters will be used for reporting the performance assessment results. The reliability of ratings is determined at the end of the rating process. When raters do not agree on the majority of ratings assigned to demonstrated

performances, additional raters would be assigned to review the same set of work samples or the entire rating process would start all over again.

Computer-aided Performance Rating Process

The computer-aided or non-traditional performance rating process operates in ways that are similar to the traditional process. Williams (2009) provided an overview of a comprehensive computer-aided process through the RosE-Portfolio System (REPS) being deployed at Rose-Hulman primarily for assessing institutional and program learning outcomes. The computer-aided rating tool is also referred to as the RosEvaluation Tool (RET).

Beginning with a set of institutional learning outcomes based on input from faculty, alumni, industry, graduate schools, and other constituents, faculty members indicate specific outcomes being addressed in their courses on a grid called a curriculum map that is updated during each academic term. Once the opportunities for developing and demonstrating relevant knowledge and skills are being identified and verified using the curriculum map, faculty members select corresponding embedded course assignments and activities in their courses that will provide the best evidence of student achievement in the outcome and direct students to submit the completed assignments to the appropriate drop boxes associated or mapped to the specific outcomes on the system. At the end of each academic year, teams of faculty evaluators or raters undergo training, using pre-defined rubrics for judging student performances for the respective outcomes. Upon completion of the performance rating process, the Office of Institutional Research, Assessment and Planning analyzes and compiles reports on rating results for supporting both institutional and program accreditation efforts by offering evidence of knowledge, skills gained and attitude changed by students through the educational process. As

for the specific steps involved in the computer-aided performance rating process, Williams (2009) identified four major steps that have been applied since 1998.

Step One

Raters, typically in a pair, review the rubric and comments, made by raters who evaluated the same outcome in previous years, associated with a specific learning outcome for evaluation. They will discuss the rubric while reviewing anchor documents, selected samples from previous rating sessions, to gain familiarity with the materials and process that help to calibrate or promote uniformity in their views when judging student performances in a more consistent way with each other and with rating teams from the past judging the same outcome.

Step Two

Raters will participate in a norming exercise, similar to the traditional setting, by independently rating a set of three identical student work samples for a specific outcome and against the pre-defined rubric. The levels of achievement for supporting institutional and program accreditation efforts are "Yes/Pass/Exemplary," "Yes/Pass," and "No/Fail" to the rating question: "Does this document meet the standard expected of a student who will graduate from Rose-Hulman." Upon completing the evaluation of the initial three identical student work samples, raters will review the results provided by the system at real-time together to identify any discrepancies in applying the rubric or other relevant issues prior to conducting the full-scale evaluation of additional submissions for the outcome.

Step Three

When the raters reach agreement in judging student performance, each rater will proceed to rate a set of unique or different submissions (typically ten) independently based on common evaluation criteria specified in the rubric. At the end of each set of unique submissions, a shared

or common submission will be assigned by the system to the raters for determining consistency in applying the rubric for judging student performance, which is typically being referred to in the educational measurement and evaluation field as inter-rater reliability. If the consistency test is positive, raters will resume judging another set of unique submissions with periodical consistency checks throughout the whole rating process. However, if raters fail to agree on the rating assigned for the common submission, the system will halt the rating process and prompt the raters to review the results and any available comments associated with the decisions together. Once an agreement is reached through matching ratings on the common submission, the system will unlock the rating sessions to allow raters to proceed with the rest of the rating process.

Step Four

Raters can provide comments as feedback on evaluated submissions and for both the performance rating administration team and raters in the coming years, so any proposed improvement suggestions can be reviewed, approved, and incorporated into the system.

Evaluation Choices of Performance Rating Process

Given the involvement of multiple raters in the performance rating process, the validity and reliability of such a process are typically evaluated through the analysis of rating agreement (Tamanini, 2008; Uebersax, 2000). Uebersax (2000) commented that in the absence of a "gold standard," the rating agreement analysis would permit certain inferences about the validity or accuracy of the given ratings as well as the rating process itself. Furthermore, the rating agreement analysis would offer insights into the reliability of ratings made by individual raters.

Depending on the chosen rating scales, common measures include Cohen's Kappa, intraclass correlation, proportion of overall agreement, raw agreement indices, G-Index (GI),

Gwet's AC_1 coefficient, and Scott/Fleiss Kappa for determining the inter-rater reliability statistics, that is, the agreement in performance ratings by different raters (Uebersax, 2000). When considering the consistency of ratings of each rater in applying the rubric for judging performances, common measures include Cohen's Kappa, G-Index (GI), Gwet's AC_1 coefficient, and intraclass correlation for determining intra-rater reliability statistics, that is, the self-consistency in performance ratings over time by each rater (Gwet, 2008b). The generic form of the rater agreement indices derives from the ratio of the overall agreement propensity between or among raters to the propensity for reaching agreement by chance. The treatment of the latter term varies among different indices in estimating the chance-agreement probability.

Due to differences in sensitivity of the statistical measures, Gwet (2008) recommended that multiple indices should be used to help estimate both the inter-rater and intra-rater reliability. Both inter-rater and intra-rater reliability statistics are important measures in determining how well the performance rating process works. They offer insights into the usefulness and validity of the process, while the number of work samples evaluated and time involved are typical factors being used for determining efficiency of the process. Hasnain and his colleagues (2004) through their experimentation with computerized decision support systems for evaluating inter-rater agreement also recommended that multiple indices to be used to address the known limitations with kappa coefficient when high trait prevalence in the performance ratings is observed.

In his study of evaluating different approaches in estimating inter-rater reliability for rating process, Stemler (2004) pointed out that the efficiency of a rating process can be achieved when raters agree on the interpretation of the rating scale and rate the performances accordingly. Once high inter-rater reliability can be ascertained, the number of work samples can be divided

among raters without having each rater to judge all of the identical work samples for a particular outcome.

Summary

Facing uncertain times, higher education institutions cannot afford to maintain the status quo, relying solely on input and output measures of the educational process to respond to the accountability demands from diverse constituents on educational outcomes. On the contrary, proactive steps should be taken by higher education institutions to examine current quality standards development efforts such as LEAP Initiative, VALUE Project, Degree Qualifications Profile and Tuning USA Initiative for establishing frameworks to define both general education and program specific learning outcomes, and the corresponding rubrics for assessing outcomes. As the frameworks evolve, direct assessments of student learning outcomes via performance rating process is likely to offer a better option for responding to questions on educational outcomes than solely relying on standardized tests, particularly through the technology-enhanced approach to minimize known obstacles associated with the typical performance rating process. Consequently, the performance rating process is more likely to be implemented more broadly within higher education institutions and across institutional boundaries to identify competencies for determining academic quality in a direct and meaningful fashion.

CHAPTER 3

METHODOLOGY

As discussed in the previous chapters, direct assessments of student learning offer relevant information for measuring and improving academic quality. Yet, the remaining challenges for adapting the performance rating process are lacking empirical evidence about the inner workings and approaches to alleviate the resource-intensive nature of the process (Palomba and Banta, 1999). This research investigated and documented the technology-enhanced, computer-aided performance rating process and assessed the feasibility for broader adaptation of such a process in an efficient manner.

Restatement of the Problem

The problem for this study is to assess the validity and reliability of a computer-aided performance rating process, which may serve as a viable option and offer relevant information for measuring and improving educational or academic quality.

As demands for evidence of student learning from higher education institutions continue to rise, so do the demands for a scalable solution that would meet both the accountability requirements and quality improvement needs. Considering that common expectations of student learning outcomes are evolving, the performance measurement process is gaining more attention due to its design for authentic assessment of student competencies. However, it is unclear if drawbacks of the traditional performance measurement process can be minimized by utilizing

available information and instructional technology, so that broader adaptation of such a process would become feasible to higher education institutions.

Restatement of Objectives

The problem for the study is addressed through three objectives:

1. Assess the validity of the computer-aided performance rating process.
2. Examine the reliability of the computer-aided performance rating process.
3. Investigate opportunities to expand the computer-aided performance rating process into a scalable solution for enhancing the efficiency of the educational outcome measurement process.

Research Design

In order to capture process data for the analyses, four raters were recruited to participate in the study. These four raters were assigned into six rating teams of two raters each to perform typical rating tasks in both the traditional and the non-traditional or computer-aided rating settings for the experiments. The traditional and the non-traditional rating tasks were both facilitated using a computer system to offer access to electronic copies of student submissions and to record ratings and comments provided by the raters. The distinct differences with the non-traditional or computer-aided performance rating process, as mentioned in the previous chapter, are as follows: rater agreements are being evaluated in real-time during the rating event and raters are not reviewing the identical set of submissions being assigned to the entire team throughout the event. The computer-aided performance rating process requires each rater on a team to evaluate independently some of the identical student submissions assigned to the entire team for monitoring rater agreement, while there are other student submissions that will only be evaluated by a single rater throughout this process.

The four raters are faculty members at Rose-Hulman Institute of Technology. Aside from having the representation of faculty members across various disciplines as a criterion for forming the teams, the curriculum map was reviewed to ensure the assigned submissions would not be coming from the raters' own courses. The rating team assignments were chosen based on each rater's past involvement in the rating process for the specific outcomes. The team selections were important for this study to allow for gathering relevant process data to estimate both the consensus between raters in each team and the self-consistency of each rater when assessing performances according to the given rubrics. The raters participated in the annual performance rating event that lasted for two days and they were compensated at a rate of $250 per day.

The outcomes selected came from the Institute Student Learning Outcomes of Rose-Hulman Institute of Technology that are mapped to the Criterion 3 of the General Criteria for Baccalaureate Level Programs specified by the Engineering Accreditation Commission (EAC) of ABET, Inc. (ABET, 2010).

At Rose-Hulman Institute of Technology, there are currently six Institutional Student Learning Outcomes with a total of 25 performance criteria:

- RH1. Leadership: Criteria A1, A2, B1, B2, and C1
- RH2. Teamwork: Criteria A1, B1, B2, and C1
- RH3. Communication: Criteria B1, B2, B3, and C1
- RH4. Cultural and Global Awareness: Criteria A1, B1, B2, B3 and C1
- RH5. Ethics: Criteria A1, B1 and C1
- RH6. Service: Criteria A1, B1, B2 and C1

Performance criteria are the measurable statements that define each learning outcome. Performance expectations are ordered with reference to Bloom's Taxonomy and reflected in the

criteria labels (A=lower-order and C=higher-order). Among these 25 performance criteria, the RH3 Communication Criterion B2 and the RH4 Cultural and Global Awareness Criterion B2 were selected for this study. The RH3 Communication Criterion B2 institutional outcome is mapped to the ABET EAC Criterion 3g. A team of two raters was assigned to conduct performance ratings of submissions for this outcome. Likewise, the RH4 Cultural and Global Awareness Criterion B2 institutional outcome is also mapped to the ABET EAC Criterion 3h. Six teams of two raters each were assigned to assess student performances under this outcome.

Work samples for these two selected outcomes were collected during academic year 2008-2009 in courses identified through the curriculum map process described in the previous chapter. There were a total of 3,095 unique student submissions collected during academic year 2008-2009. These submissions are associated with one or more of the Rose-Hulman Institutional Learning Outcomes. Among the 3,095 submissions, there were 119 unique submissions associated with the RH3 Communication Criterion B2 institutional outcome and there were 290 unique submissions associated with the RH4 Cultural and Global Awareness Criterion B2 institutional outcome. The dichotomous rating scale of "pass/fail" was used in the experiments and the corresponding question being used by the raters was, "does the document meet the standard expected of a student who will graduate from Rose-Hulman?" Aside from deriving the pass and fail percentages of the total submissions evaluated for each outcome, the respective measures for analyzing rating agreement were chosen according to the number of raters involved and the type of rating scale used for the experiments.

Estimate Process Validity

Each rater was assigned samples of student work for evaluation. The selected work samples were sorted into two groups to facilitate analyses pertaining to the research objectives.

Excluding work samples being used for the "norming" activity, the first grouping targeted work samples which the rater on a team had not rated during the annual rating event in summer of 2009; however, these work samples had been evaluated by the other rater on the same team through the computer-aided rating process during the same event. The goal for collecting additional rating results is to mimic the traditional performance rating process where both raters on each team would rate all work samples for a particular outcome. Additional identical work samples were also assigned to each team for sufficient data to be captured for the analyses. For RH3 Communication Criterion B2 as shown in Table 4, each rater on Rating Team 1 rated 39 identical work samples to emulate the traditional process with a total of 78 submissions being evaluated by the two raters.

Table 4 RH3 Communication Criterion B2 Institutional Outcome

First Grouping	Work Samples
Assigned to Rater 1 (Rated by Rater 2 in Summer 2009)	14
Assigned to Rater 2 (Rated by Rater 1 in Summer 2009)	14
Assigned to Both Rater 1 and Rater 2 (Rating Team 1)	11
Total Identical Work Samples Evaluated by Each Rater	39

For RH4 Cultural and Global Awareness Criterion B2 as shown in Table 5, each rater rated 37 identical work samples to emulate the traditional process with a total of 148 submissions being evaluated by the four raters.

Table 5 RH4 Cultural & Global Awareness Criterion B2 Institutional Outcome

First Grouping	Work Samples
Assigned to Rater 1 (Rated by Rater 2 in Summer 2009)	4
Assigned to Rater 1 (Rated by Rater 3 in Summer 2009)	4
Assigned to Rater 1 (Rated by Rater 4 in Summer 2009)	4

Table 5 (continued)

First Grouping	Work Samples
Assigned to Rater 2 (Rated by Rater 1 in Summer 2009)	4
Assigned to Rater 2 (Rated by Rater 3 in Summer 2009)	4
Assigned to Rater 2 (Rated by Rater 4 in Summer 2009)	4
Assigned to Rater 3 (Rated by Rater 1 in Summer 2009)	4
Assigned to Rater 3 (Rated by Rater 2 in Summer 2009)	4
Assigned to Rater 3 (Rated by Rater 4 in Summer 2009)	4
Assigned to Rater 4 (Rated by Rater 1 in Summer 2009)	4
Assigned to Rater 4 (Rated by Rater 2 in Summer 2009)	4
Assigned to Rater 4 (Rated by Rater 3 in Summer 2009)	4
Assigned to All Raters	13
Total Identical Work Samples Evaluated by Each Rater	37

The actual pass percentages yielded from the computer-aided (CA) process for the two outcomes in summer 2009 were compared to the pass percentages yield through this "simulated" traditional (TR) process to assess the validity or accuracy of the computer-aided process. Specifically, the pass percentage of the 77 submissions for the RH3 Communication Criterion B2 derived from the computer-aided process was compared to the pass percentage of the 78 submissions evaluated by Raters 1 and 2 as listed in Table 4. Likewise, the pass percentage of the 150 submissions for the RH4 Cultural and Global Awareness Criterion B2 derived from the computer-aided process was compared to the pass percentage of the 148 submissions evaluated by Raters 1, 2, 3 and 4 as listed in Table 5. The comparisons of pass percentages for the computer-aided (CA) and the traditional (TR) processes were conducted through the Chi-square test for homogeneity, which is equivalent to the two-proportion Z-test.

Research Objective 1: Hypothesis for Each Institutional Outcome

Ho: pCA = pTR There is no statistically significant difference in pass percentage yielded through the computer-aided (CA) performance rating process to the pass percentage yielded through the traditional (TR) performance rating process.

Ha: pCA ≠ pTR There is a difference in pass percentages yielded from the two processes.

Estimate Process Reliability

By incorporating the additional rating results captured through this study to the ones captured earlier through the computer-aided process in summer 2009, the reliability or consistency of ratings by different raters, that is, inter-rater reliability, was estimated for each outcome and each team. Specifically, proportion of overall agreement, positive agreement and negative agreement, Scott's Pi, Cohen's Kappa, G-Index, and Gwet's AC_1 coefficient were calculated for each team.

The second grouping targeted work samples that had been rated by each rater during summer 2009 for the selected outcomes as shown in Tables 6 and 7. The aim for collecting additional rating results based on second grouping of work samples is to estimate the self-consistency in assigning ratings or intra-rater reliability for each rater through repeated measurements. The same agreement indices mentioned above for estimating the inter-rater reliability were compiled for each outcome and rater.

Table 6 RH 3 Communication Criterion B2 Institutional Outcome

Second Grouping	Work Samples
Assigned to Rater 1 (Rated by Rater 1 in Summer 2009)	11
Assigned to Rater 2 (Rated by Rater 2 in Summer 2009)	11

Table 7 RH 4 Cultural and Global Awareness Criterion B2 Institutional Outcome

Second Grouping	Work Samples
Assigned to Rater 1 (Rated by Rater 1 in Summer 2009)	13
Assigned to Rater 2 (Rated by Rater 2 in Summer 2009)	13
Assigned to Rater 3 (Rated by Rater 3 in Summer 2009)	13
Assigned to Rater 4 (Rated by Rater 4 in Summer 2009)	13

Research Objective 2: For each institutional outcome being studied, the suggested interpretations of Cohen's Kappa and other kappa-like statistics, such as Scott's Pi, Gwet's AC_1 coefficient, and G-Index by Altman (1991), Bakeman & Gottman (1997), and Uebersax (2000) were referenced to assess the inter-rater reliability for each team and the intra-rater reliability for each rater respectively.

Estimate Process Efficiency

Finally, rating process efficiency was calculated for each rating setting to estimate potential benefits when using the computer-aided performance rating process.

Research Objective 3: The average minutes per submissions evaluated and the rater service fees associated with conducting performance rating of each outcome were estimated. Potential time and monetary savings when comparing the operational costs of the two performance rating processes were examined.

Data Collection

The RosEvaluation Tool (RET) was used for collecting experimental ratings assigned by each rater as well as the time spent in assessing demonstrated performances during the summer 2010 rating event. The student work samples were organized and archived on a file server. RET offered the platform for raters to examine the assigned student work samples online. The student work samples were drawn directly from the file server. At the same time, RET provided the capability for capturing and organizing the needed transactional data, such as pass or fail ratings,

written comments, indicators, etc. by each outcome and rater. The process data were stored in a database server for subsequent analyses. In order to facilitate the compilation of the intra-rater reliability statistics, relevant ratings of student work samples for the two outcomes collected during the summer 2009 rating event were extracted for the analyses.

To help ensure objectivity in data collection, training session was held prior to the performance rating event to offer information about the overall process to raters and provide an opportunity for raters to review relevant materials such as outcomes and performance criteria, primary traits and rubrics, comments from raters who evaluated submissions for the given outcomes, etc. A "norming" activity, as mentioned in the previous chapter, was required of each team to be conducted early on in the rating process to establish the common threshold in applying the corresponding rubrics when examining performances for each outcome. To facilitate the comparisons of pass percentages between the traditional and computer-aided performance rating processes, ratings collected during the norming activity and for the calibration steps when using the computer-aided process were excluded from the data analyses.

Data Analysis

The research objectives presented in Chapter 1 and restated above offer the structure for the analyses. First, the pass percentages of each outcome obtained through "simulated" traditional performance rating results were compared to the corresponding pass percentages derived from the computer-aided rating process during the summer of 2009. The hypothesis test for difference between the two pass percentages with Chi-square test was conducted for each outcome. The comparative analysis helps detect any statistically significant differences in the performance rating results and sheds light on the validity or accuracy of the given ratings associated with the computer-aided process. In light of the general acceptance of the

performance rating process that offers an authentic and direct measurement of student learning, the results of the comparative analysis add to the understanding of whether the computer-aided or non-traditional process performs in similar ways as the traditional process in assessing student performance.

Second, the reliability of the computer-aided process was examined through the derived inter-rater and intra-rater reliability indices. As mentioned earlier under the Research Design section for estimating the inter-rater reliability of the process, Scott's Pi, Cohen's Kappa, G-Index, Gwet's AC_1 coefficient and other raw agreement indices were calculated for rating teams evaluating work samples under RH3 Communication Criterion B2, and separately for the teams evaluating work samples under RH4 Cultural and Global Awareness Criterion B2. The agreement indices listed above were calculated for assessing the intra-rater reliability of the process for each outcome and rater. The accuracy or validity of the performance ratings is contingent to the reliability of the rating process. Therefore, when the computer-aided process can be shown as reliable, that is, consistent and repeatable ratings can be expected when judging specific outcomes with the corresponding rubrics, the examination of the pass percentages yielded between traditional and computer-aided processes can inform the validity study of the computer-aided process.

Finally, transactional records of the rating process were extracted and analyzed to determine the average amount of time needed for judging each work sample for each outcome. The expected time and cost estimation were derived assuming the traditional rating process for estimating transactional process efficiency between the traditional and computer-aided processes.

The followings are the key steps involved in the data collection and analysis process to address the three research objectives of this dissertation.

1. Obtained Written Approval from the Institutional Reviewer. Prior to conducting the study involving faculty members as raters during the process, the written approval from the Institutional Reviewer at Rose-Hulman Institute of Technology was obtained. The application, project description, informed consent statement, and the letter of approval are included in Appendix A.

2. Identified Representative Student Learning Outcomes. Two representative student learning outcomes that are commonly found and recognized by other higher education institutions were chosen for the study. This was discussed under Chapter 2 and the full descriptions of the selected outcomes and the corresponding rubrics are included in Appendix B.

3. Identified Work Samples for Selected Outcomes. The anchored work samples used during the summer 2009 rating event for the norming activities were extracted and random samples used during the summer 2009 rating event were selected for the research study as described under the Research Design section. Four raters who participated in the summer 2009 rating event were invited and their consents to participate in the research study were obtained.

4. Established Experimental Rating Event. The experimental rating event was setup during the summer of 2010 through the RosEvaluation Tool (RET). The screenshots of RET are presented in Appendix C.

5. Examined Experimental and Archival Data. The archived rating results from the summer 2009 rating event were examined and experimental data were analyzed.

6. Reviewed Statistical Test Results, Reliability Indices and Estimated Process Efficiency Indicators. Hypothesis testing of the pass percentages, various inter-rater and intra-rater reliability indices, average time on rating each work sample for each outcome and estimated cost

associated with the rating process were reviewed and summarized to address each research objective in Chapter 4.

7. Dissertation Committee Review. Draft copies of the dissertation were submitted to Dissertation Committee Chair and Committee members for review and comment.

Summary

The methodology and key steps involved in the research process of this dissertation were addressed in this chapter. The rationales in choosing research subjects and selected data collection and analysis methods were also discussed. The results and discussion of findings are provided in the subsequent chapters to address each research objective and the implications of the findings for adapting the computer-aided rating process for measuring academic quality.

CHAPTER 4

RESULTS

This chapter summarizes the results from examining the process outputs and the rater agreement analyses for addressing the problem of the study through the three research objectives. The results and discussions for estimating the validity, reliability and efficiency for the computer-aided performance rating process are presented below.

Estimate Process Validity

For RH3 Communication Criterion B2, the pass percentages yielded from the non-traditional or computer-aided (CA) process in summer 2009 and the "simulated" traditional (TR) process are shown in Table 8.

Table 8 RH3 Communication Criterion B2 Institutional Outcome

Performance Rating Process	Pass	Fail	Total	Pass%
Computer-Aided (CA)	56	21	77	72.7%
Traditional--Simulated (TR)	61	17	78	78.2%
Rater 1	*31*	*8*	*39*	*79.5%*
Rater 2	*30*	*9*	*39*	*76.9%*

Research Objective 1: Hypothesis for RH3 Communication Criterion B2

Ho: pCA-Comm_B2 = pTR-Comm_B2 There is no statistically significant difference in pass percentage yielded through the computer-aided performance rating process to the pass percentage yielded through the traditional performance rating process.

Ha: pCA-Comm_B2 ≠ pTR-Comm_B2 There is a difference in pass percentages yielded from the two processes.

Table 9 Cross-tabulation of Process and Rating for RH3 Communication Criterion B2

			Rating		
			Fail	Pass	Total
Process	CA	Count	21	56	77
		% within Process	27.3%	72.7%	100.0%
		% of Total	13.5%	36.1%	49.7%
	TR	Count	17	61	78
		% within Process	21.8%	78.2%	100.0%
		% of Total	11.0%	39.4%	50.3%
Total		Count	38	117	155
		% within Process	24.5%	75.5%	100.0%
		% of Total	24.5%	75.5%	100.0%

The null hypothesis (Ho) would be rejected if the p-value is less than 0.05. The conditions for conducting the Chi-Square test are having two or more independent sets of sample data, none of the expected counts is less than 1, and 20% or less of the expected counts are less than five. Since none of the expected counts is less than 1 or 5 and the two data sets are independent, the conditions for conducting the Chi-Square test were met. From the SPSS results in Table 9, the row variable, process, had two categories, so r = 2. The column variable, rating, also had two categories, so c = 2. Thus, the number of degrees of freedom was (r-1)(c-1) = (2-1)(2-1) equal to one. The critical value of $\chi 2 = 3.841$ is found from Table A.4 of Chi-Square Distribution (Triola and Franklin, 1994).

Table 10 Chi-Square Tests for RH3 Communication Criterion B2

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.628[a]	1	.428		
Continuity Correction[b]	.367	1	.545		
Likelihood Ratio	.629	1	.428		
Fisher's Exact Test				.460	.272
Linear-by-Linear Association	.624	1	.429		
N of Valid Cases	155				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 18.88.
b. Computed only for a 2x2 table

The test statistic was 0.628, which did not fall within the critical region and the p-value of 0.428 was greater than α=0.05 as shown in Table 10. Therefore, the null hypothesis was not rejected and it was concluded that the pass percentage yielded for RH3 Communication Criterion B2 through the computer-aided performance rating process did not differ from the traditional performance rating process, $\chi 2(1, N=155) = 0.628$, $p > .05$.

For RH4 Cultural and Global Awareness Criterion B2, the pass percentages yielded from the non-traditional or computer-aided process in summer 2009 and the "simulated" traditional process are shown in Table 11.

Research Objective 1: Hypothesis for RH4 Cultural and Global Awareness Criterion B2

Ho: pCA-CnGA_B2 = pTR-CnGA_B2 There is no statistically significant difference in pass percentage yielded through the computer-aided performance rating process to the pass percentage yielded through the traditional performance rating process.

Ha: pCA-CnGA_B2 ≠ pTR-CnGA_B2 There is a difference in pass percentages yielded from the two processes.

Table 11 RH4 Cultural & Global Awareness Criterion B2

Performance Rating Process	Pass	Fail	Total	Pass%
Computer-Aided (CA)	130	20	150	86.7%
Traditional--Simulated (TR)	127	21	148	85.8%
Rating Team 1 (Rater 1 and Rater 2)	*13*	*3*	*16*	*81.3%*
Rating Team 2 (Rater 1 and Rater 3)	*12*	*4*	*16*	*75.0%*
Rating Team 3 (Rater 1 and Rater 4)	*12*	*4*	*16*	*75.0%*
Rating Team 4 (Rater 2 and Rater 3)	*13*	*3*	*16*	*81.3%*
Rating Team 5 (Rater 2 and Rater 4)	*12*	*4*	*16*	*75.0%*
Rating Team 6 (Rater 3 and Rater 4)	*13*	*3*	*16*	*81.3%*
Rater 1	*13*	*0*	*13*	*100.0%*
Rater 2	*13*	*0*	*13*	*100.0%*
Rater 3	*13*	*0*	*13*	*100.0%*
Rater 4	*13*	*0*	*13*	*100.0%*

Likewise, the null hypothesis would be rejected if the p-value is less than 0.05. The conditions for conducting the Chi-Square test were met. There were two categories for each of the row and column variables as seen in Table 12, thus, the number of degrees of freedom was (r-1)(c-1) = (2-1)(2-1) equal to 1. The critical value of $\chi 2 = 3.841$ was found from Table A.4 of Chi-Square Distribution (Triola and Franklin, 1994). The test statistic was 0.046, which did not fall within the critical region and the p-value of 0.830 is greater than $\alpha=0.05$ as shown in Table 13.

Therefore, the null hypothesis was not rejected and it was concluded that the pass percentage yielded for RH4 Cultural and Global Awareness Criterion B2 through the computer-aided performance rating process did not differ from the traditional performance rating process, $\chi 2(1, N=298) = 0.046$, $p > .05$.

Table 12 Cross-tabulation of Process and Rating for RH4 Cultural & Global Awareness Criterion B2

			Rating		
			Fail	Pass	Total
Process	CA	Count	20	130	150
		% within Process	13.3%	86.7%	100.0%
		% of Total	6.7%	43.6%	50.3%
	TR	Count	21	127	148
		% within Process	14.2%	85.8%	100.0%
		% of Total	7.0%	42.6%	49.7%
Total		Count	41	257	298
		% within Process	13.8%	86.2%	100.0%
		% of Total	13.8%	86.2%	100.0%

Table 13 Chi-Square Tests for RH4 Cultural & Global Awareness Criterion B2

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.046[a]	1	.830		
Continuity Correction[b]	.002	1	.963		
Likelihood Ratio	.046	1	.830		
Fisher's Exact Test				.868	.481
Linear-by-Linear Association	.046	1	.830		
N of Valid Cases	298				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 20.36.
b. Computed only for a 2x2 table

As discussed in the previous chapter, the process validity estimation for the computer-aided process through comparing the outputs or pass percentages between the two processes offers a glimpse into whether the computer-aided process might yield significantly different pass percentages for the same outcomes being evaluated through the traditional process. Since the p-values for both hypothesis tests are greater than the significance level of .05, the null hypothesis

for each outcome of no statistically significant difference in the pass percentage yielded through the computer-aided performance rating process to the pass percentage yielded through the traditional performance rating process cannot not be rejected.

Estimate Process Reliability

For RH3 Communication Criterion B2, the rating results summary and the inter-rater indices are shown in Tables 14 and 15 respectively.

Table 14 RH3 Communication Criterion B2 Rating Results Summary

Rater 1 \ Rater 2	Pass	Fail	Total	%
Pass	29	2	31	79.5%
Fail	1	7	8	20.5%
Total	30	9	39	100.0%
%	76.9%	23.1%	100.0%	

When interpreting inter-rater reliability indices, Uebersax (2000) commented that testing the significance of the proportion of overall agreement, positive agreement and negative agreement could be done through the test of a nonzero kappa coefficient. The null hypothesis would be that raters are independent. The kappa-like indices (Scott's PI, Cohen's Kappa, G-Index and Gwet's AC_1) are all above zero as shown in Table 15, which suggest that the proportion of overall agreement, positive agreement and negative agreement are significantly different from chance.

Table 15 RH3 Communication Criterion B2 Inter-rater Reliability Indices

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.92	0.04	0.84 to 1
Positive Agreement	0.95	0.03	0.90 to 1
Negative Agreement	0.82	0.10	0.63 to 1
Scott's PI	0.77	0.12	0.52 to 1
Cohen's Kappa	0.77	0.12	0.52 to 1
G-Index	0.85	0.09	0.67 to 1

Table 15 (continued)

Method	Value	Stand Error	95% C.I.
Gwet's AC_1	0.88	0.07	0.75 to 1

Gwet (2002) provided a detailed discussion about kappa-like indices such as Scott's PI and Cohen's Kappa and he used experimental data to illustrate some limitations of these kappa-like indices when the sum of the marginal probabilities is very different from one or when high trait prevalence is observed. The kappa-like indices measure the percentage of numerical values in the main diagonal of the contingency table and then adjust these values for the amount of agreement that could be expected due to chance alone. The chance-agreement probability of each kappa-like index is estimated through a different conceptual framework. Gwet (2008) offered definitions for the G-Index and Gwet's AC_1 as alternative kappa-like indices with more robust chance-corrected statistics for evaluating the extent of agreement between raters.

For RH3 Communication Criterion B2, the high trait prevalence was observed with a high proportion of overall agreement at 0.92 together with a high positive agreement at 0.95 as seen in Table 15. Bakeman and Gottman (1997) suggested that kappa coefficient that is greater than or equal to 0.70 to be considered as the acceptable level of agreement, while Altman (1991) offered another possible interpretation of kappa coefficient:

Poor agreement = Less than 0.20

Fair agreement = 0.20 to 0.40

Moderate agreement = 0.40 to 0.60

Good agreement = 0.60 to 0.80

Very good agreement = 0.80 to 1.00

Considering the suggested interpretations of kappa coefficient by Bakeman and Gottman and Altman, the kappa values shown in Table 15 ranging from 0.77 to 0.88 indicated an acceptable level of agreement between Rater 1 and Rater 2.

When examining the self-consistency in assigning ratings, raters were assigned work samples which they had rated during summer 2009 for the RH3 Communication Criterion B2 outcome. The summary of rating results for Raters 1 and 2 are shown in Tables 16 and 18 respectively. The intra-rater reliability indices for Raters 1 and 2 are shown in Tables 17 and 19 respectively.

Table 16 RH3 Communication Criterion B2 Rating Results Summary for Rater 1

First \Second	Pass	Fail	Total	%
Pass	4	1	5	45.5%
Fail	3	3	6	54.5%
Total	7	4	11	100.0%
%	63.6%	36.4%	100.0%	

Table 17 RH3 Communication Criterion B2 Intra-rater Reliability Indices for Rater 1

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.64	0.15	0.35 to 0.92
Positive Agreement	0.67	0.16	0.36 to 0.97
Negative Agreement	0.60	0.18	0.24 to 0.96
Scott's PI	0.27	0.29	0.00 to 0.92
Cohen's Kappa	0.29	0.27	0.00 to 0.89
G-Index	0.27	0.29	0.00 to 0.92
Gwet's AC_1	0.28	0.29	0.00 to 0.93

Table 18 RH3 Communication Criterion B2 Rating Results Summary for Rater 2

First \ Second	Pass	Fail	Total	%
Pass	4	1	5	45.5%
Fail	3	3	6	54.5%
Total	7	4	11	100.0%
%	63.6%	36.4%	100.0%	

Table 19 RH3 Communication Criterion B2 Intra-rater Reliability Indices for Rater 2

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.64	0.15	0.35 to 0.92
Positive Agreement	0.67	0.16	0.36 to 0.97
Negative Agreement	0.60	0.18	0.24 to 0.96
Scott's PI	0.27	0.29	0.00 to 0.92
Cohen's Kappa	0.29	0.27	0.00 to 0.89
G-Index	0.27	0.29	0.00 to 0.92
Gwet's AC_1	0.28	0.29	0.00 to 0.93

Even though kappa-like indices (Scott's PI, Cohen's Kappa, G-Index and Gwet's AC_1) are all above zero as shown in Table 17 for Rater 1 and in Table 19 for Rater 2, the hypothesis testing of a non-zero kappa or kappa-like coefficient is not statistically significant, which suggests that the proportion of overall agreement, positive agreement and negative agreement are not necessarily different from chance for Rater1 and Rater 2.

Furthermore, all of the kappa-like indices are ranging between 0.27 to 0.29, which indicate only fair agreement or self-consistency in how Rater 1 and Rater 2 assigned ratings to their respective repeatable samples. In fact, upon closer examination of the "norming" activity when Rater 1 and Rater 2 were independently assigned ratings to three anchor documents of which they had rated during summer 2009, the rating results in Table 20 were not identical between the two sets of rating results. Specifically, the ratings for Document 2 obtained through the study were different from the rating assigned during summer 2009.

Table 20 RH3 Communication Criterion B2 Rating Results from Norming Activity

Work Sample	Pass			Fail		
	Summer 2009	Rater 1	Rater 2	Summer 2009	Rater 1	Rater 2
Document 1	1	1	1			
Document 2		1	1	1		
Document 3				1	1	1
Total	1	2	2	2	1	1

Given Rater 1 and Rater 2 did not apply the rubric in an identical way to establish the threshold in evaluating the replicates from summer 2009 during the norming activity, the intra-rater reliability indices are able to detect some inconsistencies in how each rater assigned ratings to the replicates for the study. Nevertheless, the process reliability estimation for the computer-aided process suggests that both raters have reached some consensus on how to apply the rubric, though the rater's self-consistency in assigning the ratings is only fair in this case due to some conditional changes in interpreting/applying the rubric for evaluating performances.

For RH4 Cultural and Global Awareness Criterion B2, the rating results summary and the inter-rater indices for Rating Team 1 are shown in Tables 21 and 22 respectively.

Table 21 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Team 1

Rater 1 \ Rater 2	Pass	Fail	Total	%
Pass	19	0	19	90.5%
Fail	1	1	2	9.5%
Total	20	1	21	100.0%
%	95.2%	4.8%	100.0%	

Table 22 RH4 Cultural & Global Awareness Criterion B2 Inter-rater Reliability Indices for Team 1

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.95	0.05	0.86 to 1
Positive Agreement	0.97	0.03	0.92 to 1
Negative Agreement	0.67	0.31	0.05 to 1
Scott's PI	0.64	0.33	0.00 to 1
Cohen's Kappa	0.64	0.32	0.00 to 1
G-Index	0.90	0.09	0.71 to 1
Gwet's AC_1	0.95	0.06	0.83 to 1

For Rating Team 1 and following the interpretation of inter-rater reliability indices recommended by Uebersax (2000) as discussed above, the above-zero inter-rater reliability

indices shown in Table 22 suggest that the proportion of overall agreement, positive agreement and negative agreement are significantly different from chance. Again, considering the suggested interpretations of kappa coefficient by Bakeman and Gottman and Altman, raters in Rating Team 1 have reached an acceptable level of agreement or consensus in applying the rubrics in assessing work samples for the RH4 Cultural and Global Awareness outcome.

The rating results summary and the inter-rater indices for Rating Team 2 are shown in Tables 23 and 24 respectively. For Rating Team 2, similarly, the inter-rater reliability indices shown in Table 24 suggest that the proportion of overall agreement, positive agreement and negative agreement are significantly different from chance and raters in Team 2 have reached an acceptable level of agreement or consensus in applying the rubrics in assessing work samples for the RH4 Cultural and Global Awareness outcome.

Due to the presence of low trait prevalence for fail rating, the confidence interval for the proportion of negative agreement includes a negative lower confidence limit which would be ignored.

Table 23 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Team 2

Rater 1 \ Rater 3	Pass	Fail	Total	%
Pass	18	1	19	90.5%
Fail	1	1	2	9.5%
Total	19	2	21	100.0%
%	90.5%	9.5%	100.0%	

Table 24 RH4 Cultural & Global Awareness Criterion B2 Inter-rater Reliability Indices for Team 2

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.90	0.06	0.78 to 1
Positive Agreement	0.95	0.04	0.87 to 1
Negative Agreement	0.50	0.31	-0.10 to 1
Scott's PI	0.45	0.33	0.00 to 1
Cohen's Kappa	0.45	0.33	0.00 to 1
G-Index	0.81	0.13	0.54 to 1
Gwet's AC_1	0.88	0.08	0.71 to 1

For Rating Team 3, the rating results summary is shown in Table 25. The inter-rater reliability indices shown in Table 26 suggest that the proportion of overall agreement, positive agreement and negative agreement are significantly different from chance and raters in Team 3 have reached perfect agreement or consensus in applying the rubrics in assessing work samples for the RH4 Cultural and Global Awareness outcome.

Table 25 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Team 3

Rater 1 \ Rater 4	Pass	Fail	Total	%
Pass	19	0	19	90.5%
Fail	0	2	2	9.5%
Total	19	2	21	100.0%
%	90.5%	9.5%	100.0%	

Table 26 RH4 Cultural & Global Awareness Criterion B2 Inter-rater Reliability Indices for Team 3

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	1.00	0.00	1.00 to 1
Positive Agreement	1.00	0.00	1.00 to 1
Negative Agreement	1.00	0.00	1.00 to 1
Scott's PI	1.00	0.00	1.00 to 1
Cohen's Kappa	1.00	0.00	1.00 to 1

Table 26 (continued)

Method	Value	Stand Error	95% C.I.
G-Index	1.00	0.00	1.00 to 1
Gwet's AC_1	1.00	0.00	1.00 to 1

The rating results summary for Rating Team 4 is provided in Table 27. Given the high trait prevalence was observed in Table 28 with a high proportion of overall agreement at 0.86 together with a high positive agreement at 0.92, the more robust G-Index and Gwet's AC_1 were considered when estimating the inter-rater reliability. Under such condition, both Scott's PI and Cohen's Kappa indices were deficient in measuring agreement between raters 2 and 3 in Team 4, which yielded negative or extremely low coefficient values when one would expect the extent of agreement between the two raters would be higher. The inter-rater reliability indices in the table suggest that the proportion of overall agreement and positive agreement are significantly different from chance and raters in Rating Team 4 have reached consensus in applying the rubrics in assessing work samples for the RH4 Cultural and Global Awareness outcome.

Table 27 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Team 4

Rater 2 \ Rater 3	Pass	Fail	Total	%
Pass	18	2	20	95.2%
Fail	1	0	1	4.8%
Total	19	2	21	100.0%
%	90.5%	9.5%	100.0%	

Table 28 RH4 Cultural & Global Awareness Criterion B2 Inter-rater Reliability Indices for Team 4

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.86	0.08	0.71 to 1
Positive Agreement	0.92	0.04	0.84 to 1

Table 28 (continued)

Method	Value	Stand Error	95% C.I.
Negative Agreement	0.00	0.00	0.00 to 0
Scott's PI	-0.08	0.04	0 to 0.02
Cohen's Kappa	-0.07	0.05	0 to 0.03
G-Index	0.84	0.10	0.63 to 1
Gwet's AC_1	0.71	0.15	0.40 to 1

For Rating Team 5, the rating results summary and the inter-rater reliability indices are shown in Tables 29 and 30 respectively. The inter-rater reliability indices in Table 30 suggest that the proportion of overall agreement, positive agreement and negative agreement are significantly different from chance and raters in Rating Team 5 have reached an acceptable level of agreement or consensus in applying the rubrics in assessing work samples for the RH4 Cultural and Global Awareness outcome. Again, due to the presence of low trait prevalence for fail rating, the confidence interval for the proportion of negative agreement includes a negative lower confidence limit which would be ignored.

Table 29 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Team 5

Rater 2 \ Rater 4	Pass	Fail	Total	%
Pass	18	0	18	85.7%
Fail	2	1	3	14.3%
Total	20	1	21	100.0%
%	95.2%	4.8%	100.0%	

Table 30 RH4 Cultural & Global Awareness Criterion B2 Inter-rater Reliability Indices for Team 5

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.90	0.06	0.78 to 1
Positive Agreement	0.95	0.04	0.87 to 1
Negative Agreement	0.50	0.31	-0.10 to 1

Table 30 (continued)

Method	Value	Stand Error	95% C.I.
Scott's PI	0.45	0.33	0.00 to 1
Cohen's Kappa	0.46	0.31	0.00 to 1
G-Index	0.81	0.13	0.54 to 1
Gwet's AC_1	0.88	0.08	0.71 to 1

The rating results summary for Rating Team 6 is provided in Table 31. The inter-rater reliability indices shown in Table 32 suggest that the proportion of overall agreement, positive agreement and negative agreement are significantly different from chance and raters in Rating Team 6 have reached an acceptable level of agreement or consensus in applying the rubrics in assessing work samples for the RH4 Cultural and Global Awareness outcome.

Table 31RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Team 6

Rater 3 \ Rater 4	Pass	Fail	Total	%
Pass	19	0	19	90.5%
Fail	1	1	2	9.5%
Total	20	1	21	100.0%
%	95.2%	4.8%	100.0%	

Table 32 RH4 Cultural & Global Awareness Criterion B2 Inter-rater Reliability Indices for Team 6

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	0.95	0.05	0.86 to 1
Positive Agreement	0.97	0.03	0.92 to 1
Negative Agreement	0.67	0.31	0.05 to 1
Scott's PI	0.64	0.33	0.00 to 1
Cohen's Kappa	0.64	0.32	0.00 to 1
G-Index	0.90	0.09	0.71 to 1
Gwet's AC_1	0.95	0.06	0.83 to 1

When examining the self-consistency in assigning ratings for the RH4 Cultural and Global Awareness outcome, raters were assigned work samples of which they had rated during summer 2009. The summary of rating results and the intra-rater reliability indices for Rater 1 are shown in Tables 33 and 34 respectively.

Table 33 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Rater 1

First \Second	Pass	Fail	Total	%
Pass	13	0	13	100.0%
Fail	0	0	0	0.0%
Total	13	0	13	100.0%
%	100.0%	0.0%	100.0%	

Table 34 RH4 Cultural & Global Awareness Criterion B2 Intra-rater Reliability Indices for Rater 1

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	1.00	0.00	1.00 to 1
Positive Agreement	1.00	0.00	1.00 to 1
Negative Agreement	-	-	-
Scott's PI	-	-	-
Cohen's Kappa	-	-	-
G-Index	1.00	0.00	1.00 to 1
Gwet's AC_1	1.00	0.00	1.00 to 1

Due to the high trait prevalence observed in Table 33 and with a high proportion of overall agreement at 1.00 together with a high positive agreement at 1.00 identified in Table 34, the G-Index and Gwet's AC_1 were used for estimating the intra-rater reliability. Since both G-Index and Gwet's AC_1 are all above zero, these indices suggest that the proportion of overall agreement and the positive agreement are significantly different from chance. Moreover, both indices are equal to 1.0, which indicate perfect agreement or self-consistency in how Rater 1 assigned ratings to their respective repeatable samples associated with the RH4 Cultural and

Global Awareness outcome. The same observations were recorded for Raters 2, 3 and 4 who were involved in evaluating work samples for the same outcome. The respective rating results summaries and intra-rater reliability indices for Raters 2, 3, and 4 are shown in Tables 35 through 40 for each rater.

Table 35 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Rater 2

First \Second	Pass	Fail	Total	%
Pass	13	0	13	100.0%
Fail	0	0	0	0.0%
Total	13	0	13	100.0%
%	100.0%	0.0%	100.0%	

Table 36 RH4 Cultural & Global Awareness Criterion B2 Intra-rater Reliability Indices for Rater 2

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	1.00	0.00	1.00 to 1
Positive Agreement	1.00	0.00	1.00 to 1
Negative Agreement	-	-	-
Scott's PI	-	-	-
Cohen's Kappa	-	-	-
G-Index	1.00	0.00	1.00 to 1
Gwet's AC_1	1.00	0.00	1.00 to 1

Table 37 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Rater 3

First \Second	Pass	Fail	Total	%
Pass	13	0	13	100.0%
Fail	0	0	0	0.0%
Total	13	0	13	100.0%
%	100.0%	0.0%	100.0%	

Table 38 RH4 Cultural & Global Awareness Criterion B2 Intra-rater Reliability Indices for Rater 3

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	1.00	0.00	1.00 to 1
Positive Agreement	1.00	0.00	1.00 to 1
Negative Agreement	-	-	-
Scott's PI	-	-	-
Cohen's Kappa	-	-	-
G-Index	1.00	0.00	1.00 to 1
Gwet's AC_1	1.00	0.00	1.00 to 1

Table 39 RH4 Cultural & Global Awareness Criterion B2 Performance Rating Results Summary for Rater 4

First \Second	Pass	Fail	Total	%
Pass	13	0	13	100.0%
Fail	0	0	0	0.0%
Total	13	0	13	100.0%
%	100.0%	0.0%	100.0%	

Table 40 RH4 Cultural & Global Awareness Criterion B2 Intra-rater Reliability Indices for Rater 4

Method	Value	Stand Error	95% C.I.
Proportion of Overall Agreement	1.00	0.00	1.00 to 1
Positive Agreement	1.00	0.00	1.00 to 1
Negative Agreement	-	-	-
Scott's PI	-	-	-
Cohen's Kappa	-	-	-
G-Index	1.00	0.00	1.00 to 1
Gwet's AC_1	1.00	0.00	1.00 to 1

Perfect agreement or self-consistency in applying rubric to evaluate work samples associated with the RH4 Cultural and Global Awareness Criterion 2 was observed for each of the four raters as shown in Tables 4.28 through 4.33 above. These raters have also applied the rubric

in an identical way to establish the threshold for the study when evaluating the anchor documents from summer 2009 during the norming activity. The intra-rater reliability indices are offering supporting evidence to these observations through the examination of how each rater assigned ratings to the replicates for the study.

Overall, the process reliability estimation for the computer-aided process suggests that all raters have reached consensus in how to apply the rubric, as well as maintained fair to good self-consistency in assigning the ratings according to the rubric for evaluating performances for the two outcomes.

Estimate Process Efficiency

Finally, rating process efficiency was calculated for each rating setting to estimate potential benefits when using the computer-aided performance rating process. For RH3 Communication Criterion B2, the time spent on evaluating work samples by each rater is presented in Table 41. For RH4 Cultural and Global Awareness Criterion B2, the time spent on evaluating work samples by each rater is presented in Table 42.

Table 41 RH3 Communication Criterion B2 Descriptive Statistics on Time Spent in Rating (Excluding "Norming" Activity)

Each Sample	Min	Max	Mean	Standard Deviation
Rater 1	< 1 min.	8 mins.	2.5 mins.	1 min.
Rater 2	< 1 min.	7 mins.	2.5 mins.	1 min.
Mean			2.5 mins.	

Table 42 RH4 Cultural & Global Awareness Criterion B2 Descriptive Statistics on Time Spent in Rating (Excluding "Norming" Activity)

Each Sample	Min	Max	Mean	Standard Deviation
Rater 1	< 1 min.	7 mins.	1 mins.	1 min.
Rater 2	< 1 min.	2 mins.	1 mins.	< 1 min.
Rater 3	< 1 min.	8 mins.	1 mins.	1 min.
Rater 4	< 1 min.	7 mins.	2 mins.	1 min.
Mean			1.3 mins.	

Excluding work samples being used for the "norming" activity and calibration steps involved in the computer-aided process, there were 77 unique work samples for the RH3 Communication Criterion 2 being evaluated with the computer-aided or non-traditional process during summer 2009. The estimated time for completing the rating tasks with the computer-aided and traditional processes using the mean minute per sample from Table 41 are shown in Table 43. For RH4 Cultural and Global Awareness Criterion B2, there were 150 unique work samples, excluding those being used for the "norming" activity and calibration steps, evaluated using the computer-aided process during summer 2009. The estimated time for completing the rating tasks with the computer-aided and traditional processes using the mean minute per sample from Table 42 are presented in Table 44.

Table 43 RH3 Communication Criterion B2 Estimated Time Spent in Rating (Excluding "Norming" Activity and Calibration Steps)

Performance Rating Process	Sample	Review per Sample	Mean (Minute/Sample)	Estimated Time Spent
Computer-Aided (CA)	77	1	2.5 mins.	193 mins.
Traditional (TR)	77	2	2.5 mins.	385 mins.

Table 44 RH4 Cultural & Global Awareness Criterion B2 Estimated Time Spent in Rating (Excluding "Norming" Activity and Calibration Steps)

Performance Rating Process	Sample	Review per Sample	Mean (Minute/Sample)	Estimated Time Spent
Computer-Aided (CA)	150	1	1.3 mins.	195 mins.
Traditional (TR)	150	2	1.3 mins.	390 mins.

Considering the current daily compensation amount for each rater is $250, the estimated hourly pay rate is about $38.50 for approximately 6.5 hours of performance rating work excluding lunch and other breaks. As suggested by Stemler (2004) and discussed in the previous chapter, the efficiency of the rating process can be enhanced assuming raters agree on the interpretation of the rubric and evaluate the demonstrated performances accordingly. Provided that high inter-rater reliability is garnered, the number of work samples can be divided among raters without involving two or more raters in judging the performance of each identical work sample for a particular outcome.

For this study, the process validity estimation offers support that the computer-aided or non-traditional performance rating process yields similar outputs or pass percentages as with the traditional performance rating process. In addition, the process reliability estimation points toward an acceptable level of rater agreement obtained for each rating team. With the basic conditions are being met, the process efficiency estimation in terms of time and cost savings for the performance rating process are derived and illustrated in Table 45. Through division of labor, the estimated time and money spent in assessing student performance could be cut half through the computer-aided rating process with two-rater teams.

Table 45 Process Efficiency Estimation (Time and Cost Savings)

Outcome	Time			Cost		
	TR	CA	Savings	TR	CA	Savings
RH3 Communication Criterion B2	385 mins.	193 mins.	192 mins	$247.04	$123.52	$123.52
RH4 Cultural & Global Awareness Criterion B2	390 mins.	195 mins.	195 mins	$250.25	$125.13	$125.12

Summary

With the known challenges for adapting performance rating process for direct assessments of student learning, this study aims at addressing specific issues related to performance rating process validity, reliability and efficiency. This chapter has offered descriptions about the inner workings of a typical performance rating process and the key differences between the traditional and the non-traditional, or computer-aided, processes. In addition, process validity and reliability of the computer-aided rating process have been estimated through experiments using two learning outcomes that are common to colleges and universities across the country. Rating results obtained from the experiments through four raters, which formed six two-rater teams, offer evidence of how a non-traditional or computer-aided process can be utilized to evaluate performances in a similar fashion to the traditional rating process.

Finally, process efficiency measures in terms of potential savings of time and labor cost associated with the computer-aided process have been compiled to estimate gains of using such process with two-rater teams. A summary of the study, further discussions of the research findings, other considerations relevant to the study, conclusion, and recommendations for future research are presented in Chapter 5.

CHAPTER 5

DISCUSSION

This chapter presents a summary of the pilot study, discussion of the findings associated with each of the three research objectives (process validity, process reliability and process efficiency), other considerations related to the research design, conclusion of the study, and the recommendations for future research.

Summary

From what has been discussed in the previous chapters on the rising demands for evidence of student learning and ways to improve academic quality, it seems as though these demands will not be going away anytime soon. The recurring theme associated with the higher education's iron triangle—access, cost, and quality—can be found when scanning the media headlines on higher education. The global economic crisis continues to hamper the ability of higher education institutions to respond to challenges of providing wider access to educational opportunities in light of rising costs and higher expectations of academic quality (Daniel, Kanwar & Uvalic-Trumbic, 2009).

Although standardized tests offer one way for higher education institutions to provide comparable evidence of student learning primarily at the institutional level, this quality measure has its limitations in terms of satisfying the perennial demands for actionable information in making curriculum improvement changes and understanding what students actually can do

particularly in their program areas (Koretz, 2008.) On the other hand, traditional performance rating process as a quality measure offers a relatively more flexible and direct way for higher education institutions to define, measure and demonstrate student learning in line with their program and institutional goals and objectives. However, the scalability of such process and the costs, such as labor and time, associated with it could render the performance rating process less practical or affordable for large-scale implementation by higher education institutions. This study described the inner workings of a computer-aided performance rating process and examined the validity and reliability of such non-traditional performance rating process. The purpose of this dissertation is to determine whether the computer-aided process can serve as a viable and scalable option that would minimize the known drawbacks of the traditional performance rating process.

Through the literature review process, using direct assessment of student learning as an academic quality measure was introduced and the needs for empirical research into performance rating process and solutions for addressing the known challenges associated with such process were identified. The process validity associated with the computer-aided process was estimated through the comparisons of the outputs from the computer-aided and the traditional processes using the Chi-square test for homogeneity. The results of the hypothesis tests indicated that the computer-aided process could produce similar outputs as with the traditional process. Moreover, three rater agreement tests and four kappa-like rater agreement indices were used to estimate the process reliability associated with the computer-aided process. When the norming activity is followed as with the rating teams on Cultural and Global Awareness B2 , the results of the rater agreement analyses showed that the overall agreements between the two raters in each team (inter-rater reliability) and of the same rater over time (intra-rater reliability) were within

the commonly acceptable range for the measures, that is, consistent performance rating results could be expected of the computer-aided process. Furthermore, the estimated differences in terms of time spent on the performance rating process and costs associated with the process were contrasted between the computer-aided and the traditional settings. The results suggested that efficiency could be gained through the utilization of the computer-aided process. In addition to the review of the strengths and weaknesses associated with the computer-aided performance rating process, opportunities for process improvement are discussed in this chapter.

Discussion of the Findings

As discussed in earlier chapters, the performance rating process has a long history and in the form of "juried reviews," its root is dated back as early as 1063 (Bresciani, 2006). Demonstrating the acquisition of knowledge and skills by oral and written examinations are common and longstanding practices inside and outside of classrooms in higher education. Subject matter experts assess the level of competence of learners according to the demonstrated performances and the associated rubrics. This quality measure is best suited when the performing tasks are multidimensional and complex, such as those involving higher order thinking skills of analysis, evaluation and synthesis (Bloom, 1956). In fact, if expected competencies could be easily codified, the standardized tests would likely be a more cost-effective measure to determine competency achievement of learners when considering costs and labor associated with the performance rating process. Thus lies the dilemma that prompts for the perennial question of what students learn at higher education institutions among diverse stakeholders that often extends beyond general education outcomes and into the specific program areas. So far, the standardized test results and students' self-reported gains through surveys are not able to satisfy the needs of employers, legislatures, professional societies and the general

public for such information. With the understanding that the performance rating process has the potential for capturing and offering the needed information on student competencies, this study focused on the known and significant challenges associated with such process and explored the feasibility of utilizing a technology-enhanced process that could improve the efficiency in assessing student performance.

In order to serve as a viable academic quality measure, the technology-enhanced or computer-aided performance rating process must possess similar characteristics as with the traditional process, of which the computer-aided process should yield similar performance ratings when judging demonstrated performances under a comparable operational environment. The validity of the computer-aided process is approximated through the examination of the accuracy of the performance ratings against those yielded through the traditional process. In addition, the computer-aided process should yield consistent and expected results when applying the same rubrics and under the same operational condition. The reliability of the computer-aided process is estimated through the examination of inter-rater and intra-rater reliability indices. Furthermore, the computer-aided process should enhance the overall efficiency of the performance rating process, given that the resource, time and labor-intensive nature of the performance rating process is a major deterrent to the use of such direct measurement process. The efficiency of the computer-aided process is projected by comparing the expected time and cost estimates between the computer-aided and traditional performance rating processes.

Estimate Process Validity

As noted in the previous chapter, the pass percentages in assessing student performances by using the computer-aided process are comparable to those yielded through the traditional process for both the RH3 Communication Criterion B2 and the RH4 Cultural and Global

Awareness Criterion B2. Aside from using a typical two-rater team for assessing student performance for the RH3 Communication Criterion B2, six different pairings of four raters were chosen to share the workload and derive the pass percentages for the RH4 Cultural and Global Awareness Criterion B2 for analysis. The results obtained are encouraging news, which indicates that comparable performance ratings can be expected of the computer-aided process and flexibility can be built in when more raters would be needed for evaluating larger quantity of student submissions.

Estimate Process Reliability

Reliability indices offer a glimpse of the variability in the performance rating process due to differences in the application of rubrics between raters and self-consistency in the interpretation and subsequently the application of the rubrics by each rater. Such examination is useful in determining how reliable the performance rating process is, as well as, offering directions for making specific improvement changes to the process. For RH3 Communication Criterion B2, although the reliability indices indicated relatively good agreement between Rater 1 and Rater 2 in judging performance and the yielded performance ratings were not likely coming through random guesses, they pointed out only a fair agreement in how Rater 1 and Rater 2 applied the rubrics individually and over time. The fair intra-rater agreement triggered an audit into the norming activity involved prior to the starting of the rating event and it is clear that there was some difference in the interpretation of the rubrics by Rater 1 and Rater 2 for the experiment.

For RH4 Cultural and Global Awareness Criterion B2, the reliability indices for both the inter-rater reliability for all six rating teams and the intra-rater reliability for all four raters suggested that the agreements among the four raters were high in applying the rubrics when

judging performances and self-consistency in applying the rubrics was maintained by each of the four raters. When considering improvement changes using the derived reliability indices, they point to the needs for revisiting the norming activity for raters involved in judging submissions under the RH3 Communication Criterion B2 and reviewing the corresponding rubrics to clarify performance goals and expectations.

When trait prevalence exists among the observed performance ratings as with the case for RH4 Cultural and Global Awareness Criterion B2, it becomes problematic, if not impossible, to derive the traditional kappa-like reliability indices. In that situation, the more robust reliability indices such as G-Index and Gwet's AC_1 offer alternative ways to estimate the extent of agreement between raters.

Estimate Process Efficiency

Efficiency measure is only meaningful for consideration when the process itself is regarded valid and reliable. Based on the results for estimating the validity and reliability of the computer-aided performance rating process in this study, the various indicators suggest that the computer-aided process is comparable to the traditional process. Specifically with situations where high inter-rater reliability can be ascertained as discussed by Stemler (2004), the computer-aided process could be employed to improve the efficiency for applying the performance rating process. Aside from using the computer-aided process to streamline the logistics for supporting the performance rating activities, efficiency can be further enhanced by division of labor while inter-rater reliability is being monitored in real-time. The computer-aided process enables higher education institutions to explore the fit and feasibility of applying such direct assessment method for obtaining relevant information for making curriculum improvements, as well as, fine-tuning the assessment process itself.

Other Considerations

Even though this study has chosen two typical student learning outcomes that can be found in many general education outcome statements posted by higher education institutions across the country, the performing rating process is by no mean restricted to serve as a quality measure at the institutional level. As discussed in the earlier chapters, program-specific learning outcomes such as civil engineering and even course learning outcomes can also utilize the same process for assessing student competencies in specific program areas and for individual courses, particularly for those with multiple class sections. Again, given that measuring academic quality is a multidimensional challenge, it is not to say that the performance rating process alone can answer all of the questions from diverse constituents about what students learn in college. There are times that other quality measures such as standardized tests or student surveys could serve as effective means in addressing specific questions related to student learning. However, the performance rating process, particularly in the computer-aided mode, can minimize the logistical challenges for conducting direct assessments of student learning and provide information that may inform both curriculum and quality process improvement changes.

For this study, a dichotomous rating scale is chosen for judging the quality of performance that satisfies the competence expectations at the institutional level. Two-rater teams are used to provide the direct assessment of student work samples. The corresponding reliability indices are chosen to help measuring the extent of agreement between raters accordingly. For other measurement settings such as program level assessment, other rating scales may apply. The corresponding rating scales may be polytomous, that is ordered category or purely nominal in nature, or continuous. At the same time, multiple raters may be needed to assess student performance. In these cases, the reliability indices covered in the earlier chapters can be extended

to estimate rater agreement by using non-dichotomous rating scales and by having three or more raters involved in the performance rating process. When continuous rating scale is chosen, additional parametric methods can be applied to estimate rater agreement such as intraclass correlations and related Analysis of Variance models. Although various statistical methods for estimating the validity and reliability of the performance rating process are available for evaluating the quality of ratings derived from such process, it is important to first consider the explicit goals for the performance rating tasks and select the appropriate rating scales and number of raters to garner the desirable data for informing decisions. For quality assurance purposes and measuring student competencies at the institutional level, a dichotomous rating scale such as pass or fail with the typical pairing of raters may be sufficient to determine if the performance goals are attained. As for program or course level performance measurement, the specificity of expected performances could vary and so as the choice for identifying the appropriate rating scales and number of raters for measuring the desired traits. For example, ABET accredited programs are expected to categorize and provide student work samples from courses into three quality groups—high, medium, low. If the performance rating process were applied in this case, the appropriate rating scale that satisfies this condition would be polytomous. Another related consideration for program level measurement, in terms of efficiency of using performance rating process as a quality measure, is to utilize a curriculum map as discussed in earlier chapters to identify and coordinate data collection points and efforts. One of the ways to achieve greater efficiency is to consider using targeted course-embedded assignments that may inform improvement decisions on refining course, program and institutional outcomes. In light of the current condition of the global economy and considering the ongoing trend of funding for higher education, strategic alignment of student outcome

assessment activities should be beneficial to higher education institutions to better allocate available resources for achieving the performance goals and responding to the question on educational outcomes.

As reflected in the analysis of both inter-rater and intra-rater reliability indices for the RH3 Communication Criterion B2, the importance of having carefully constructed performance criteria and the associated rubrics are critical for successfully applying the performance rating process and yielding helpful information for making improvement changes. Raters will face challenges in interpreting and applying the rubrics in the same manner for judging performance when the competence expectations are not clearly articulated. Furthermore, the norming activity for establishing a common threshold in applying the rubrics for specific learning outcomes also plays an important role and it will have an influence to the outcome of the performance rating process.

Last but not least, it is helpful to place the emphasis back on measuring academic quality through the lens of student learning outcomes assessment, given that teaching and learning are the crux of the education process. Having an assessment plan and schedule to chart out the expected outcomes, types of quality measures and frequency of conducting each assessment appropriate for the level of measurement, such as course, program and institutional, should render the overall academic quality measurement process in a more consistent and manageable fashion. It is true that establishing a clear and robust framework and articulating competence expectations according to the missions of the higher education institutions are no simple tasks; however, meaningful feedback is more likely to be acquired through such deliberate efforts in articulating the performance goals, aligning resources, and identifying the appropriate means to achieve them. Ultimately, it is impossible to determine the right course of actions or make any

improvement changes without measurement. Keeping in mind the saying often attributed to Albert Einstein, "Not everything that can be counted counts, and not everything that counts can be counted," or a similar phrase put together by Elliot Eisner (2005), "Not everything that matters is measurable, and not everything that is measurable matters," serves as a helpful reminder that careful considerations should be made with regard to the practical importance of defining the key and relevant performance goals and choosing the appropriate measures for the tasks at hand.

Conclusion

This pilot study centered on the idea that one of the best measures of academic quality as suggested in the literature is through direct assessment of student learning, given the importance of student learning in the missions and purposes of all higher education institutions. The research study responded to the call for empirical research into the performance rating process, offered descriptions of how such process would operate in academic settings, and examined the validity, reliability and efficiency of a computer-aided process. Through the examination of the above three aspects of the computer-aided process, this study suggested that the computer-aided performance rating process would warrant a closer look by members of the educational community on the possibility of adapting such process that could minimize the major and known drawbacks of direct assessment of student learning via the traditional performance rating process.

With no end in sight for the ongoing financial crisis that restricts resources available to educational services, a viable and scalable performance measurement solution would be advantages for higher education institutions that seek to respond to the perennial question raised

by diverse constituent groups on what students learn at the institutions while gathering relevant information for making curricular improvement changes.

Recommendations for Future Research

This study is one of the first to compare a computer-aided performance rating process to the traditional one and to assess the validity and reliability of the computer-aided process. It serves to fill a gap in our current understanding of the role of performance rating process for measuring academic quality and the feasibility of applying available information and instructional technology to mitigate the major drawbacks associated with such process. There are numerous opportunities for further research of adaptability and sustainability of the computer-aided performance rating process. The current study has obtained positive indicators from the experiments that the computer-aided process possesses similar operational characteristics as the traditional process.

Further research might include replicating the experiments locally using a different batch of student work samples for the same outcomes, as well as, selecting different student learning outcomes to determine if the computer-aided process itself would continue to perform in a similar fashion. Moreover, similar experiment could be replicated at other higher education institutions to examine the transferability of the computer-aided performance rating process. Additional study might estimate process reliability using polytomous rating scales and different sizes of rating teams using the computer-aided rating process to determine the scalability of such process under different sets of operating conditions.

Given the establishment of common threshold in applying the rubrics during the rater training stage is important to the reliability of the overall rating process, it is recommended that further research to be conducted in determining the optimal settings for the norming activity,

such as the numbers and types of samples to be evaluated to help raters develop and maintain a common basis for reaching consensus in judging performance.

Finally, with the proliferation of information and instructional technology and the advancement of statistical methods, it is highly recommended that the latest development associated with computer-aided performance rating process and reliability indices within the higher education community and in the fields of human performance technology, medicine and psychology should be monitored and reviewed. New opportunities may be discovered and innovative solutions may then be adapted for measuring academic quality in a more effective and efficient way.

REFERENCES

ABET. (n.d.a). *About ABET*. Retrieved May 23, 2011, from http://www.abet.org/about-abet/

ABET. (n.d.b). *Assessment Planning Resources.* Retrieved May 23, 2011, from http://www.abet.org/assessment-planning-resources/

ABET 2010. *Criteria for accrediting engineering programs.* Baltimore, MD: ABET, Inc. Retrieved May 23, 2011, from http://www.abet.org/Linked%20Documents-UPDATE/Program%20Docs/abet-eac-criteria-2011-2012.pdf

Adelman, C. (2009). *The bologna process for U.S. eyes: Re-learning higher education in the age of convergence.* Washington, DC: Institute for Higher Education Policy.

Alsmadi, A. (2005). Assessing the quality of students' ratings of faculty members at Mu'tah University. *Social Behavior and Personality*, *33*(2), pp. 183-188.

Altman, D.G. (1991). *Practical statistics for medical research.* London: Chapman and Hall.

American Society of Civil Engineers. (n.d.). *About ASCE.* Retrieved June 9, 2010, from http://www.asce.org/About-ASCE/

American Society of Civil Engineers. (2008). *Civil engineering body of knowledge for the 21st century: Preparing the civil engineer for the future* (2nd ed.). Reston, VA: American Society of Civil Engineers. Retrieved June 9, 2010, from http://www.asce.org/CE-Body-of-Knowledge/

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses.* Chicago, IL: University of Chicago Press.

Association of American Colleges and Universities. (n.d.a). *LEAP: Liberal education and America's promise.* Retrieved June 9, 2010, from http://www.aacu.org/leap/documents/EssentialOutcomes_Chart.pdf

Association of American Colleges and Universities. (n.d.b). *VALUE: Valid assessment of learning in undergraduate education.* Retrieved June 9, 2010, from http://www.aacu.org/value/index.cfm

Association of American Colleges and Universities. (2005). *Liberal education outcomes: A preliminary report on student achievement in college.* Washington, DC: Author. Retrieved June 9, 2010 from http://www.aacu.org/advocacy/pdfs/LEAP_Report_FINAL.pdf

Association of American Colleges and Universities. (2011). *AAC&U statement on the Lumina Foundation for Education's proposed degree qualifications profile.* Washington, DC: Author. Retrieved March 1, 2011, from http://www.aacu.org/about/statements/documents/lumina_dqs_2011.pdf

Bakeman, R., & Gottman, J.M. (1997). Assessing observer agreement. In, *Observing interaction: An introduction to sequential analysis* (2nd ed., pp. 56-80). New York: Cambridge University Press.

Berenson, M.L., Levine, D.M. & Krehbiel, T.C. (2012). *Basic Business Statistics*, (12th ed.). Upper Saddle River, NJ: Prentice Hall Press.

Blackmur, D. (2008). A critical analysis of the INQAAHE guidelines of good practice for higher education quality assurance agencies. *Higher Education: The International Journal of Higher Education and Educational Planning*. *56*(6), pp. 723-734.

Bloom, B.S. (1956). *A taxonomy of educational objectives.* Chicago, IL: University of Chicago Press.

Bresciani, M.J. (2006). *Outcomes-based academic and co-curricular program review: a compilation of institutional good practices*. Sterling, VA: Stylus.

Brown, J.D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies*, *22*(2), pp. 91-139.

Bucher, J. (2008, Fall). The basics of a quality calibration program. *CMM Quarterly*, pp. 13-15.

Buu, Y.A. (2003). *Statistical analysis of rater effects* (Doctoral dissertation). Retrieved March 12, 2012, from http://etd.fcla.edu/UF/UFE0001244/buu_y.pdf

Carmines, E. G. & Zeller, R.A. (1991). *Reliability and validity assessment.* Newbury Park: Sage Publications.

Carpenter, C.L., Davis, M.J., Hafter, J.C., Harbaugh, J.D., Hertz, R., Johnson, Jr., E.C., Jones, M., Kloppenberg, L.A., Perez, T.E., Pierce, R.C., Schmoke, K.L., and Worthen, K.J. (2008, July 27). *Report of the outcome measures committee*. Retrieved June 9, 2010, from http://apps.americanbar.org/legaled/committees/subcomm/Outcome%20Measures%20Final%20Report.pdf

College of DuPage. (n.d.). *Comparing the AQIP and Baldrige criteria*. Retrieved May 24, 2010, from http://www.cod.edu/Academic/AcadServ/AQIP/chart4.pdf

Center for Higher Education Policy Studies (CHEPS) 2010. The first decade of working on the European Higher Education Area. The Bologna Process independent assessment executive summary, overview and conclusions. Retrieved November 1, 2010, from http://www.ond.vlaanderen.be/hogeronderwijs/bologna/2010_conference/documents/IndependentAssessment_executive_summary_overview_conclusions.pdf

Chun, M. (2002). Looking where the light is better: A review of the literature on assessing higher education quality. *Peer Review*, *4*(2/3), 16-25.

Council for Higher Education Accreditation. (2011). *Roles and relationships: Accreditation and the federal government.* Retrieved March 1, 2011, from http://www.chea.org/pdf/NACIQI_Feb_2011.pdf

Daniel, J., Kanwar, A., & Uvalic-Trumbic, S. (2009, March-April). Breaking higher education's iron triangle: access, cost and quality. *Change*, *41*, 30-35.

Davies, A. & Le Mahieu, P. (2003). Assessment for learning: reconsidering portfolios and research evidence. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Innovation and Change in Professional Education: Optimising new modes of assessment: In search of qualities and standards* (pp. 141-169). Dordrecht: Kluwer Academic Publishers.

Deaton, J.E., Bell, B., Fowlkes, J., Bowers, C., Jentsch, F., and Bell, M.A. (2007). Enhancing team training and performance with automated performance assessment tools. *The International Journal of Aviation Psychology*. *17*(4), pp. 317-331.

Dill, D.D., and Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher Education*, *49*(4): pp. 495-533.

Eisner, E. (2005, September). Back to whole. *Educational Leadership*, *63*(1), pp. 14-18.

Ewell, P.T. (2008). Assessment and accountability in America today: Background and context. *New Directions for Institutional Research*, *2008*(S1), pp. 7-17.

Evan, J.R., and Dean, J.W. (2003). *Total quality: Management, organization, and strategy.* Mason, Ohio: South-Western College Publishing.

Evans, A. (2003). *Reliability.* Retrieved March 12, 2012, from http://www.cchil.org/cru/images/education/df4469473b14ed0f33dc48efed4fd740.pdf

Fink, A., ed. (1995). *How to measure survey reliability and validity.* (7). Thousand Oaks, CA: Sage.

Gillen, A., Bennett, D., & Vedder, R. (2010). *The inmates running the asylum? An analysis of higher education accreditation.* Washington, DC: Center for College Affordability and Productivity.

Goldie, J., Schwartz, L., McConnachie, A., Jolly, B. & Morrison, J. (2004). Can students' reasons for choosing set answers to ethical vignettes be reliably rated? Development and testing of a method. *Medical Teacher*. *26*(8), pp. 713-718.

Gwet, K.L. (2002). *Kappa statistic is not satisfactory for assessing the extent of agreement between raters*. Retrieved May 1, 2011, from http://www.agreestat.com/research_papers.html

Gwet, K.L. (2008a). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*. *61*(1), 29-48.

Gwet, K.L. (2008b). Intrarater reliability. *Wiley Encyclopedia of Clinical Trials*. Retrieved May 1, 2011, from http://www.agreestat.com/research_papers.html

Hacker, A., & Dreifus, C. (2010). *Higher education? How colleges are wasting our money and failing our kids—and what we can do about it.* New York, NY: Times Books.

Hasnain, M., Onishi, H. & Elstein, A.S. (2004). Inter-rater agreement in judging errors in diagnostic reasoning. *Medical Education. 38*(6), pp. 609-616.

Higher Learning Commission. (2003). *Handbook of accreditation.* Retrieved March 1, 2011, from https://content.springcm.com/content/DownloadDocuments.ashx?Selection=Document%2C10611003%3B&accountId=5968

Hunt Jr., J.B., & Tierney, T. (2006). *American higher education: how does it measure up for the 21st century.* San Jose, CA: The National Center for Public Policy and Higher Education.

Illinois State Board of Education. (1995). *Assessment handbook: A guide for developing assessment programs in Illinois schools.* Springfield, IL: Illinois State Board of Education

Institute for Evidence-Based Change. (2010). *Tuning educational structures: a guide to the process version 1.0.* Retrieved March 1, 2011, from http://tuningusa.org/TuningUSA/b7/b70c4e0d-30d5-4d0d-ba75-e29c52c11815.pdf

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kuo, W. (2006). Assessment for U.S. engineering programs. *IEEE Transactions on Reliability*, *55*(1), pp. 1-6.

Leskes, A. (2002). Beyond confusion: An assessment glossary. *Peer Review*, *4*(2/3), pp. 42-43.

Leuven/Louvain-la-Neuve Communiqué. (2009). Retrieved March 1, 2011, from http://www.ehea.info/Uploads/Declarations/Leuven_Louvain-la-Neuve_Communiqu%C3%A9_April_2009.pdf

Lewis, R.C., and Booms, B.H. (1983). The marketing aspects of service quality. In *Emerging Perspectives on Services Marketing* (L. Berry, G. Shostack and G. Upah, eds), pp. 99-104. Chicago: American Marketing Association.

Lewis, R.G. and Smith, D.H. (1994). *Total quality in higher education.* Delray Beach, FL: St. Lucie Press.

Linn, R.L. and Miller, M.D. (2009). *Measurement and assessment in teaching,* (10th ed.). Upper Saddle River, NJ: Prentice Hall Press.

Lombardi, M.M. (2008). Making the grade: the role of assessment in authentic learning. *EDUCAUSE Learning Initiative.* Washington, DC: EDUCAUSE.

Lumina Foundation. (2011). *The degree qualifications profile.* Retrieved March 1, 2011 from http://www.luminafoundation.org/publications/The_Degree_Qualifications_Profile.pdf

Maguad, B.A. (2009). Using QFD to integrate the voice of the customer into the academic planning process. *Proceedings of ASBBS Annual Conference*: Las Vegas, *16*(1).

Mazor, K.M., Zanetti, M., Alper, E.J., Hatem, D., Barrett, S.V., Meterko, V., Gammon, W. & Pugnaire, M.P. (2007). Assessing professionalism in the context of an objective structured clinical examination: an in-depth study of the rating process. *Medical Education*. *41*(4), pp. 331-340.

McMartin, F., McKenna, A., & Youssefi, K. (2000). Scenario assignments as assessment tools for undergraduate engineering education. *IEEE Transactions on Education*, *43*(2), pp. 111-119.

Mueller, J. (2011). *Mueller's glossary of authentic assessment terms*. Retrieved March 12, 2012, from: http://jfmueller.faculty.noctrl.edu/toolbox/glossary.htm

Newton, J. (2000, July). Feeding the beast or improving quality?: Academics' perceptions of quality assurance and quality monitoring. *Quality in Higher Education*, *6*(2), pp. 153-164.

Ottolini, M.C., Cuzzi, S., Tender, J., Coddington, D.A., Focht, C., Patel, K.M., & Greenberg, L. (2007). Decreasing variability in faculty ratings of student case presentations: A faculty development intervention focusing on reflective practice. *Teaching and Learning in Medicine*. *19*(3), pp. 239-243.

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

Prados, J.W., Peterson, G.D., & Lattuca, L.R. (2005). Quality assurance of engineering education through accreditation: The impact of engineering criteria 2000 and its global influence. *Journal of Engineering Education*, *94*(1), pp. 165-184.

Roch, S.G., Paquin, A.R., & Littlejohn, T.W. (2009). Do raters agree more on observable items. *Human Performance*, *22*(5), pp. 391-409.

Rogers, G.M., & Chow, T. (2000). Electronic portfolios and the assessment of student learning. *Assessment Update*, *12*(1), 4-6.

Rogers, G.M., & Williams, J.M. (2001). Promise and pitfalls of electronic portfolios: lessons learned from experience. *A Collection of Papers on Self-Study and Institutional Improvement.* Chicago, IL: North Central Association of Colleges and Schools.

Rollins, A.M. (2011). *A case study: Application of the balanced scorecard in higher education* (Doctoral dissertation). Retrieved March 12, 2012, from http://sdsu-dspace.calstate.edu/xmlui/bitstream/handle/10211.10/1382/Rollins_Andrea.pdf?sequence=1

Rothchild, M.T. (2011). *Accountability mechanisms in public multi-campus systems of higher education* (Doctoral dissertation). Retrieved March 12, 2012, from ProQuest Dissertations and Theses. (UMI No. 3449162)

Saunders, V.M. (2007). *Does the accreditation process affect program quality? A qualitative study of perceptions of the higher education accountability system on learning* (Doctoral dissertation). Retrieved March 12, 2012, from ProQuest Dissertations and Theses. (UMI No. 1394669371)

Schamber, J.F. and Mahoney, S.L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *The Journal of General Education*, *55*(2), pp. 103 -137.

Shah, M., & Brown, G. (2009). The rise of private higher education in Australia: Maintaining quality outcomes and future challenges. *Proceedings of the Australian Universities Quality Forum* (AUQF), 138-143. Melbourne: Australian Universities Quality Agency

Shavelson, R. (2007 January/February). Assessing student learning responsibility: From history to an audacious proposal. *Change*, *39*(1), pp. 27-33.

Shulman, L.S. (2007 January/February). Counting and recounting: Assessment and the quest for accountability. *Change*, *39*(1), pp. 20-25.

State Higher Education Executive Officers. (2005). *Accountability for better results—A national imperative for higher education.* Retrieved March 1, 2011, from http://www.sheeo.org/account/accountability.pdf

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved May 1, 2011, from http://PAREonline.net/getvn.asp?v=9&n=4

Stiggins, R. J. (1987). The design and development of performance assessments. Educational *Measurement: Issues and Practice*, *6*, pp. 33-42.

Sullivan, B.F. & Thomas, S.L. (2007). Documenting student learning outcomes through a research-intensive senior capstone experience: Brining the data together to demonstrate progress. *North American Journal of Psychology*, *9*(2), pp. 321-330.

Tamanini, K.B. (2008). *Evaluating differential rater functioning in performance ratings: Using a goal-based approach* (Doctoral dissertation). Retrieved March 12, 2012, from: http://etd.ohiolink.edu/send-pdf.cgi/Tamanini%20Kevin%20B.pdf?ohiou1222283256

Thomson, G., and Douglass, J.A. (2009). *Decoding learning gains: Measuring outcomes and the pivotal role of the major and student backgrounds.* Center for Studies in Higher Education, University of California, Berkeley. Retrieved March 12, 2012, from http://cshe.berkeley.edu/publications/docs/ROPS-GT-JD-Decoding-5-30-09.pdf

Tsung, F., Li, Y., & Jin, M. (2008). Statistical process control for multistage manufacturing and service operations: a review and some extensions. *International Journal of Services Operations and Informatics*, *3*(2), 191-204.

Triola, M. F., and Franklin, L.A. (1994). *Business statistics: understanding populations and processes*. Reading, MA: Addison-Wesley.

Uebersax, J. (May 2000). Rater and Diagnostic Test Agreement. *Latent Structure Analysis.* Retrieved May 1, 2011, http://www.john-uebersax.com/stat/

University & College Accountability Network (U-CAN). (n.d.). *About U-CAN*. Washington, DC: National Association of Independent Colleges and Universities. Retrieved March 1, 2011, from http://www.ucan-network.org/about-u-can

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Retrieved March 1, 2011, from http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

Vaughn, J. (2002). Accreditation, commercial rankings, and new approaches to assessing the quality of university research and education programmes in the United States. *Higher Education in Europe*, *27*(4), pp. 433-441.

Voluntary System of Accountability Program. (n.d.). *About VSA*. Retrieved March 1, 2011, from http://www.voluntarysystem.org/index.cfm?page=about_vsa

Walvoord, B. (2004). *Assessment clear and simple: A practical guide for institutions, departments, and general education.* San Francisco, CA: Jossey-Bass.

Weldy, T.G. and Turnipseed, D.L. (2010). Assessing and improving learning in business schools: Direct and indirect measures of learning. *Journal of Education for Business*, *85*, pp. 268-273.

Wergin, J.F. (2005 May/June). Higher education: Waking up to the importance of accreditation. *Change*, *37*(3), pp. 35-41.

Williams, J.M. (2009). Institutional uses of rubrics and e-portfolios: Spelman College and Rose-Hulman Institute. *Peer Review*, *11*(1), 24-26.

Williams, J.M. (2010, March). Evaluating what students know: using the RosE Portfolio System for institutional and program outcomes assessment tutorial. *IEEE Transactions on Professional Communication*, *53*(1), 46-57.

Yaffee, R.A. (2003). *Common correlation and reliability analysis with SPSS for Windows*. Retrieved March 12, 2012, from http://www.nyu.edu/its/statistics/Docs/correlate.html

Zeithaml, V.A., Parasuraman, A., & Berry, L.L. (1990). *Delivering quality service: Balancing customer perceptions and expectations.* New York, NY: The Free Press.

APPENDIX A: CERTIFICATION OF EXEMPTION (IR# RHS0135)

Application for Review of Research Involving Human Participants and Related Documents

ROSE-HULMAN INSTITUTE OF TECHNOLOGY 5500 WABASH AVENUE TERRE HAUTE, IN 47803-3920 PHONE: 812-877-8228 FAX: 812-877-8895 www.rose-hulman.edu
ELECTRICAL AND COMPUTER ENGINEERING

4/10/2012

Timothy Chow
CM 11
Rose-Hulman Institute of Technology

RE: An Experimental Examination of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunites
IR# RHS0135

Dear Timothy,

I have reviewed your proposed extension of the study listed above, pursuant to Rose-Hulman Institute of Technology's Human Research Protection Policy and 45 CFR 46. The extension is granted through 4/10/2013. You will need to inform me of the completion of the research or report the need for a continuation beyond one year. Should you need to make modifications to your protocol or informed consent forms that do not fall within the exemption categories, you must receive approval prior to modification.

Any problems involving risk to participants or others, injury or other adverse effects experienced by participants, and incidents of noncompliance must be reported to the IR via phone or e-mail immediately.

If you have any questions, please contact me. I wish you well on completing your study.

Sincerely,

Bob Throne
IR

x8414 throne@rose-hulman.edu

ROSE-HULMAN INSTITUTE OF TECHNOLOGY 5500 WABASH AVENUE TERRE HAUTE, IN 47803-3920 PHONE: 812-877-8105 FAX: 812-877-8895 www.rose-hulman.edu
ELECTRICAL AND COMPUTER ENGINEERING

5/3/2011

Timothy Chow
CM 11
Rose-Hulman Institute of Technology

RE: An Experimental Examination of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunites
IR# RHS0135

Dear Timothy,

I have reviewed your proposed extension of the study listed above, pursuant to Rose-Hulman Institute of Technology's Human Research Protection Policy and 45 CFR 46. The extension is granted through 5/2/12. You will need to inform me of the completion of the research or report the need for a continuation beyond one year. Should you need to make modifications to your protocol or informed consent forms that do not fall within the exemption categories, you must receive approval prior to modification.

Any problems involving risk to participants or others, injury or other adverse effects experienced by participants, and incidents of noncompliance must be reported to the IR via phone or e-mail immediately.

If you have any questions, please contact me. I wish you well on completing your study.

Sincerely,

Bob Throne
IR

x8414 throne@rose-hulman.edu

ROSE-HULMAN
INSTITUTE OF TECHNOLOGY
Department of Electrical & Computer Engineering

5/27/10

Timothy Chow
CM 11
Rose-Hulman Institute of Technology

RE: An Experimental Examination of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunites
IR# RHS0135

Dear Timothy,

I have reviewed your proposed study listed above, pursuant to Rose-Hulman Institute of Technology's Human Research Protection Policy and 45 CFR 46. This proposed study falls within an exempt category (2) and is therefore considered exempt from Institutional Review Board review. You will need to inform me of the completion of the research or report the need for a continuation beyond one year. Should you need to make modifications to your protocol or informed consent forms that do not fall within the exemption categories, you must receive approval prior to modification.

Informed Consent: Approved

Any problems involving risk to participants or others, injury or other adverse effects experienced by participants, and incidents of noncompliance must be reported to the IR via phone or e-mail immediately.

If you have any questions, please contact me. I wish you well on completing your study.

Sincerely,

Robert D. Throne

Robert Throne
IR

X8414 throne@rose-hulman.edu

5500 WABASH AVENUE • TERRE HAUTE, INDIANA 47803-3920
PHONE: 812-877-8228 • FAX: 812-877-8895
http://www.rose-hulman.edu

Rose-Hulman Institute of Technology

Application for Review of Research Involving Human Participants

For IR Use Only
IR File No: _____
Date received: _____
Approval expires: _____

Federal regulations and Rose-Hulman Institute of Technology's Human Research Protection Policy require that all research involving humans as subjects be reviewed and approved prior to the commencement of recruitment and data collection. Any person (RHIT faculty member, student, staff member, or other person) wanting to engage in human subject research must receive written approval from the Institutional Reviewer (IR) or, if required, by the Institutional Review Board (IRB) before conducting the research.

1. Title of Project: An Experimental Examination of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunities

2. Principal Investigator:
   ☐ Faculty ☐ Student* X Staff ☐ Other—specify _____
   *Students are required to have a faculty, staff, or professional sponsor.
   Campus Box No. or Mailing Address: CM 11
   Phone: 8910 Email: chow@rose-hulman.edu

3. Co-Investigator(s) or Sponsor (student research must be sponsored by faculty or qualified staff):
   Include all additional investigators with contact information.
   ☐ Faculty ☐ Student ☐ Staff ☐ Other—specify _____
   Campus Box No. or Mailing Address:

   Phone: Email:

4. Project Description: Provide a brief description using layperson's terms of the proposed research, including purpose and research questions or hypothesis. Describe briefly how information will be collected, recorded, stored, and disseminated and procedures for maintaining confidentiality. List any funding sources sought or attained. Describe incentives, if any, being offered for participation in the study and any costs, if any, to the participants.

See attached.

5. Informed Consent: State below or attach your informed consent form/statement.

See attached.

6. Indicate the categories of participants to be included in the study (check all that apply):
   ☐ Abortuses/Fetuses ☐ Patients
   ☐ Decisionally Impaired ☐ Prisoners
   ☐ Decisionally Impaired (Institutionalized) ☐ Pregnant Women
   ☐ Minors (17 years of age or less, give age range:____ X Students
   X Normal Volunteers

7. Does this research involve information that may identify participants? X Yes ☐ No

8. Describe the informed consent procedures to be followed, including circumstances under which consent will be sought and obtained, who will seek it, and the method for documenting consent.

See attached.

9. Risks:

X The risks are minimal (i.e. the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests).
☐ The risks are greater than minimal.

10. Some categories of research may be exempt from full IRB review, including those below. Check the categories that apply to your research project:

☐ 1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

X 2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless (i) information obtained is recorded in such a manner that human participants can be identified, directly or through identifiers linked to the participants; and (ii) any disclosure of human participants' responses outside the research could reasonably place the participants at risk of criminal or civil liability or be damaging to the participants' financial standing, employability, or reputation. *Note: According to 45 CFR 46.401, if the participants are children, this exemption applies only to research involving educational tests or observations of public behavior when the investigator(s) does not participate in the activities being observed.*

☐ 3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior that is not exempt under #2 (above) of this section if: (i) the human participants are elected or appointed public officials or candidates for public office; or (ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

☐ 4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that participants cannot be identified, directly or through identifiers linked to the participants.

☐ 5. Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.

☐ 6. Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed or (ii) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture.

11. Estimated starting date: June 1, 2010 Estimated completion date: December 31, 2010

INVESTIGATOR ASSURANCE

I certify that the information provided for this project is correct and that no other procedures will be used in this protocol. I agree to conduct this research as described. I will request approval from the IR for changes to the study's protocol and/or consent procedures and will not implement the changes until I receive approval for these changes. I understand that changes may require approval of the IRB before proceeding. I will comply with the RHIT's Human Research Protection Policy for the conduct of ethical research. I will report significant or adverse effects or noncompliance to the IR via phone or e-mail immediately, and then in writing within 5 days of occurrence. I will be responsible for ensuring that the work of colleagues involved with this project complies with this protocol. I will complete, on request by the IR, a Continuation Request or Completion of Research Activities forms.

Principal Investigator's Signature — Date: 5/26/10

Faculty/Qualified Staff Sponsor — Date

Submit all materials to the IR: Robert Throne, Campus Mail #114 x8414

Rose-Hulman Institute of Technology

Application for Review of Research Involving Human Participants

For IR Use Only
IR File No: _____
Date received: _____
Approval expires: _____

Federal regulations and Rose-Hulman Institute of Technology's Human Research Protection Policy require that all research involving humans as subjects be reviewed and approved prior to the commencement of recruitment and data collection. Any person (RHIT faculty member, student, staff member, or other person) wanting to engage in human subject research must receive written approval from the Institutional Reviewer (IR) or, if required, by the Institutional Review Board (IRB) before conducting the research.

1. Title of Project: An Experimental Examination of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunities

4. Project Description: Provide a brief description using layperson's terms of the proposed research, including purpose and research questions or hypothesis. Describe briefly how information will be collected, recorded, stored, and disseminated and procedures for maintaining confidentiality. List any funding sources sought or attained. Describe incentives, if any, being offered for participation in the study and any costs, if any, to the participants.

With the known shortfalls with standardized examination (such as artificial time limit for problem-solving and relevance of topics to what's being taught in the classroom) to serve as the primary measure of education quality, the "alternative" criterion-referenced, rubric-based measurement approach is a promising complement or potentially a replacement for the standardized examinations. This is particularly true for judging fundamental knowledge and skills expected of the general education of postsecondary students in any college or university across the nation. The primary criticism of the criterion-referenced, rubric-based measurement approach is the lack of empirical evidence of expanding the typically known as "subjective" and "time and labor intensive" measurement practice into an "objective" and "effective/efficient" measurement of educational quality through demonstrated outcomes. The proposed research is attempted to shed insights into such expansion of "criterion-referenced, rubric-based" measurement method for assessing education quality beyond classroom level by focusing on the reliability of a computer-aided measurement process (this computer-aided process is co-developed in-house by me and others here in our organization.) The questions to be investigated through this process validation research are:

- Is the computer-aided performance measurement (rating) method (not a computer-grader like GRE Essay Grader to use computer to rate or grade but to use computer algorithm to monitor the rating process and identify events where deviation in applying scoring rubrics in a uniform manner among raters in a team) comparable to the generally accepted traditional performance measurement method using a team of evaluators/raters to judge performance quality independently and use aggregate results in making claims of such performance?

- Examine inter-rater reliability (Cohen's kappa) or temporal consistency in performance measurement: perform statistical analysis on experimental data to compare and address reliability of this alternative performance measurement process.

- Examine scalable option through computer-aided method: perform statistical analysis on experimental data to address effectiveness/efficiency of the process (primary criticism of this measurement approach): estimate time/effort differences assuming inter-rater reliability to be sustained from above.

Experimental design will be applied to analyze data that is currently stored in a database on a server administered and monitored by the Office of Instructional, Administrative and Information Technology (with restricted access only by authorized personnel) and new data to be gathered through an upcoming performance rating session. The new data will be stored in the same database as described above. The server is protected by our Institute's firewall and using Kerberos authentication protocol to restrict further access to the server.

The primary participants of this study will be raters (faculty members) who evaluate performances demonstrated through student submissions/work samples. There is no compensation for participating in this study. The data collection process will be a part of the upcoming institutional rating session.

Any information obtained during this study will be kept strictly confidential. The data will be stored on Institute's servers behind firewall. Participants and individual information will not be identified in the dissertation or in any relevant publications/presentations.

5. Informed Consent: State below or attach your informed consent form/statement.

See attached.

8. Describe the informed consent procedures to be followed, including circumstances under which consent will be sought and obtained, who will seek it, and the method for documenting consent.

Raters will be presented the above Informed Consent Form prior to conducting the experiment. The Informed Consent Form will provide participants information on the purpose of the research, procedures, risks and/or discomforts, benefits, confidentiality, compensation, opportunity to ask questions, freedom to withdraw, and right. A copy of the signed consent form will be given to each participant and original copy will be kept by the investigator.

ROSE-HULMAN INSTITUTE OF TECHNOLOGY 5500 WABASH AVENUE TERRE HAUTE, IN 47803 PHONE: 812-877-1511 FAX: 812-877-8362 WWW.ROSE-HULMAN.EDU

June 1, 2010

An Experimental Examination of a Computer-Aided Performance Rating Process and the Associated Process Improvement Opportunities

You are being invited to participate in a research study about evaluating a "criterion-referenced, rubric-based" performance measurement (rating) method for assessing education quality beyond classroom level by focusing on the reliability of a computer-aided performance rating process. The strengths and weaknesses associated with the computer-aided rating process will be explored for informing future process quality improvement changes. This study is being conducted by Timothy Chow, from the Office of Institutional Research, Planning and Assessment at Rose-Hulman Institute of Technology. This study is being conducted as part of a dissertation.

You were selected as a possible participant in this study because of your prior experience and involvement in using the computer-aided performance rating process, which would enable the investigator to establish some level of control for the experimental examination. There are no known risks if you decide to participate in this research study. There are no costs to you for participating in the study. The (performance rating and related) information you provide will be used for comparing the traditional and computer-aided performance rating processes, examining temporal consistency (inter-rater reliability) in performance rating process and estimating differences in resource requirements for these rating processes. The data collection process (reviewing and rating student submissions) for this research will take approximately 60-90 minutes to complete as part of the Institute Rating event. The information collected may not benefit you directly, but the information learned in this study should provide more general benefits.

Any information obtained during this study which could identify you will be kept strictly confidential. The data will be stored behind Institute's firewall on servers administered and monitored by IAIT and will only be seen by the investigator and other authorized supporting personnel during the study. The Institutional Review Board may inspect these records. Should the data be published, no individual information will be disclosed.

Your participation in this study is voluntary. By proceeding to review and rate student submissions identified for this research, you are voluntarily agreeing to participate. You are free to decline to participate in this study for any reason and at any time without adversely affecting your relationship with the investigator. Your decision will not result in any loss or benefits to which you are otherwise entitled.

If you have any questions about the study, please contact Timothy Chow, via campus mail to CM11, by phone at 8910, or by e-mail to chow@rose-hulman.edu.

If you have any questions about your rights as a research subject or if you feel you've been placed at risk, you may contact the Institute Reviewer at Rose-Hulman, Professor Robert Throne, via campus mail to CM 114, by phone at 8414, or by e-mail at throne@rose-hulman.edu.

Principal Investigator: Timothy Chow, CM 11, Tel: 8910, Fax: 8931, e-mail: chow@rose-hulman.edu

APPENDIX B: SELECTED OUTCOMES AND RUBRICS

Selected Rose-Hulman Institutional Student Learning Outcomes and Rubrics

RH 3. Communication, regardless of the media, requires unique skills whether communicating with individuals or with groups.

Criterion B2. Adapt technical information for a non-specialized audience.
Primary traits: A passing submission for this criterion must:
Address technical problems or concepts.
Appear free of unexplained technical jargon and acronyms.

Potential documents: Documents appropriate for this criterion include (but are not limited to): An outreach presentation/activity teaching science, mathematics, or engineering content to K-12 students; a description of current research in science, mathematics, or engineering written as if for submission to a popular press magazine or newspaper; an oral presentation to individuals skilled in disciplines other than the technical discipline of the subject matter.

Additional information:
1. We define technical problems or concepts as related to science, math or engineering. We would classify "economics" as a science. Note that the type of science (social, physical, biological, etc.) is not specified.

RH 4. Cultural and Global Awareness requires perception and understanding of the cultural perspectives and social systems that define human communities.

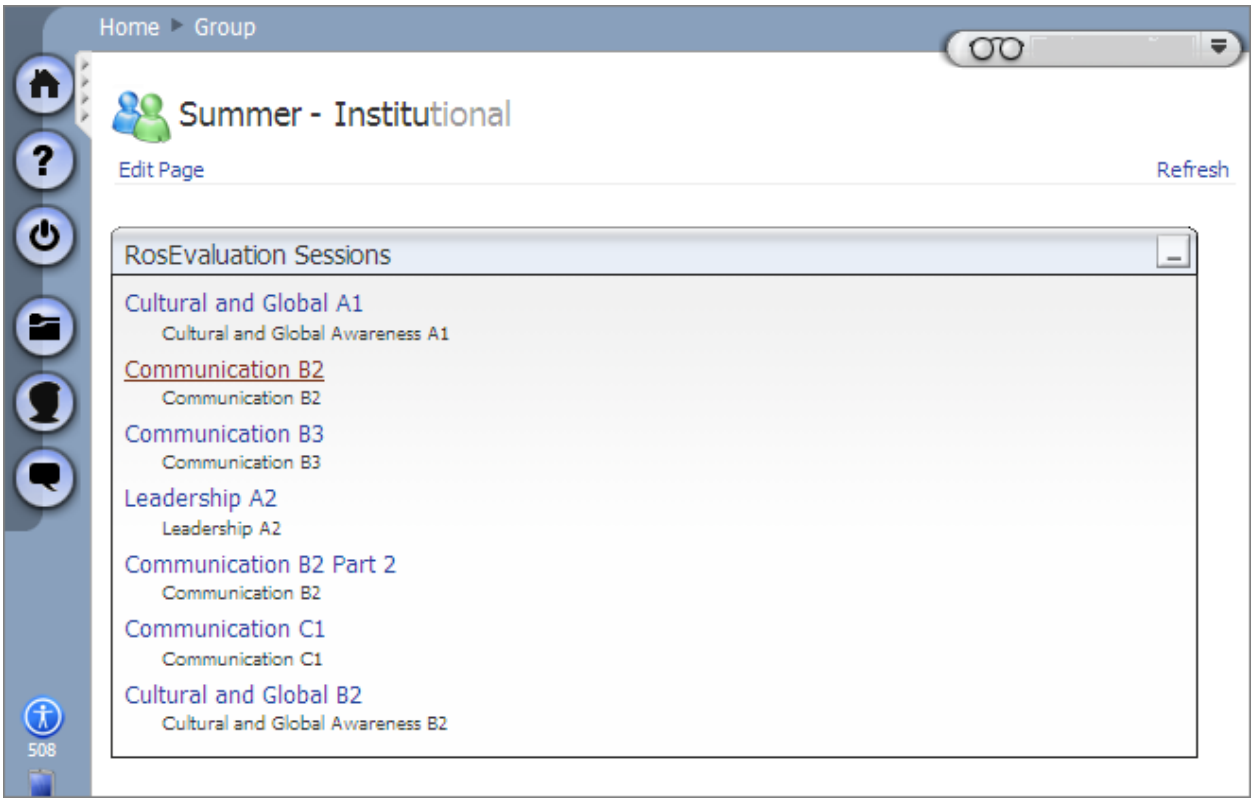
Criterion B2. Analyze beliefs, backgrounds, cultures, or societies different from your own.
Primary traits: A passing submission for this criterion must:
1. Analyze, interpret, or evaluate aspects of a non-US culture/society, or of multiple cultures/societies (one of which may be from US).
2. Sustain an international and/or comparative perspective throughout.

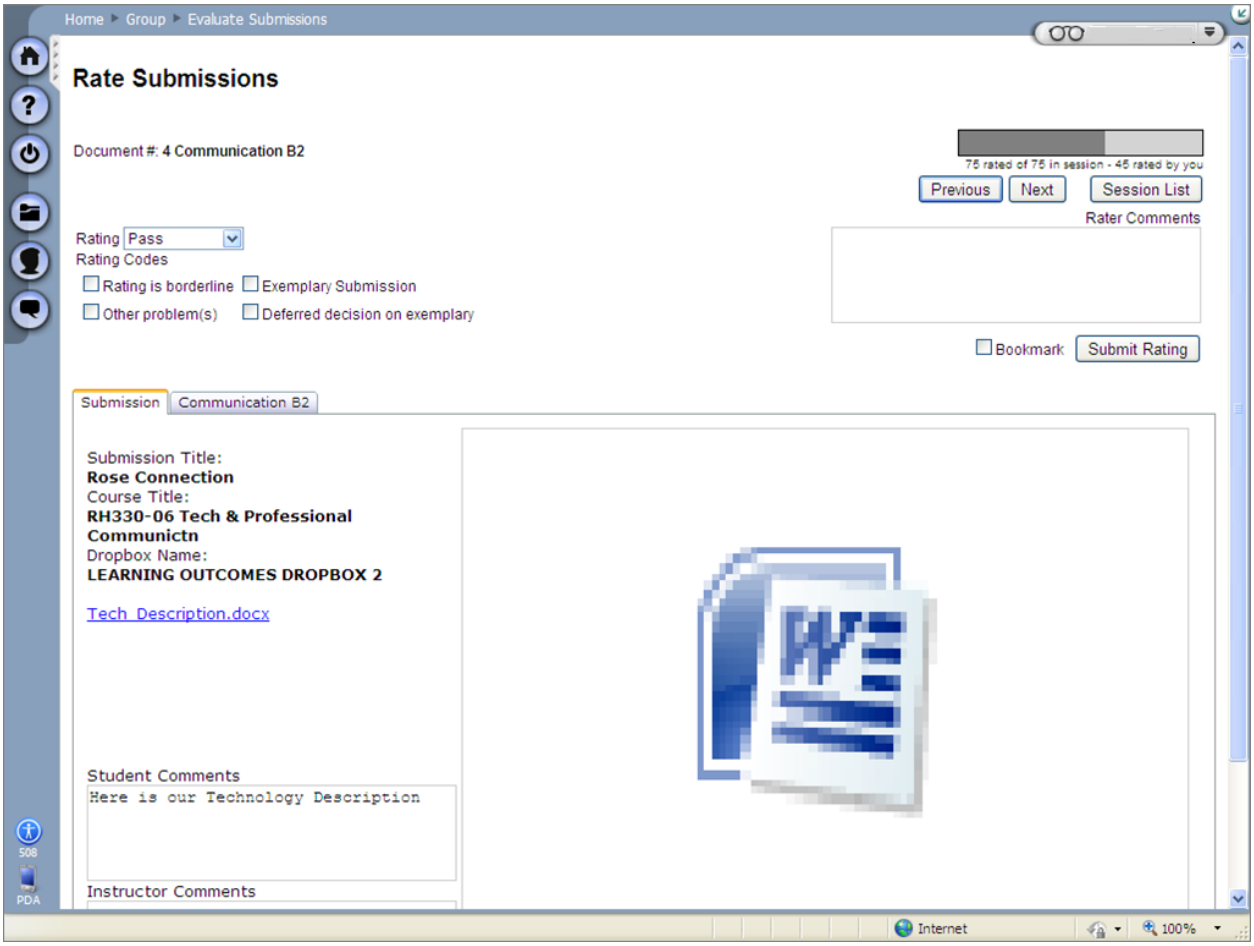
Potential documents: Documents for this criterion may be drawn primarily from the HSS GLOBAL STUDIES category, courses whose primary focus is on the examinations of other societies or interrelationships among multiple societies.

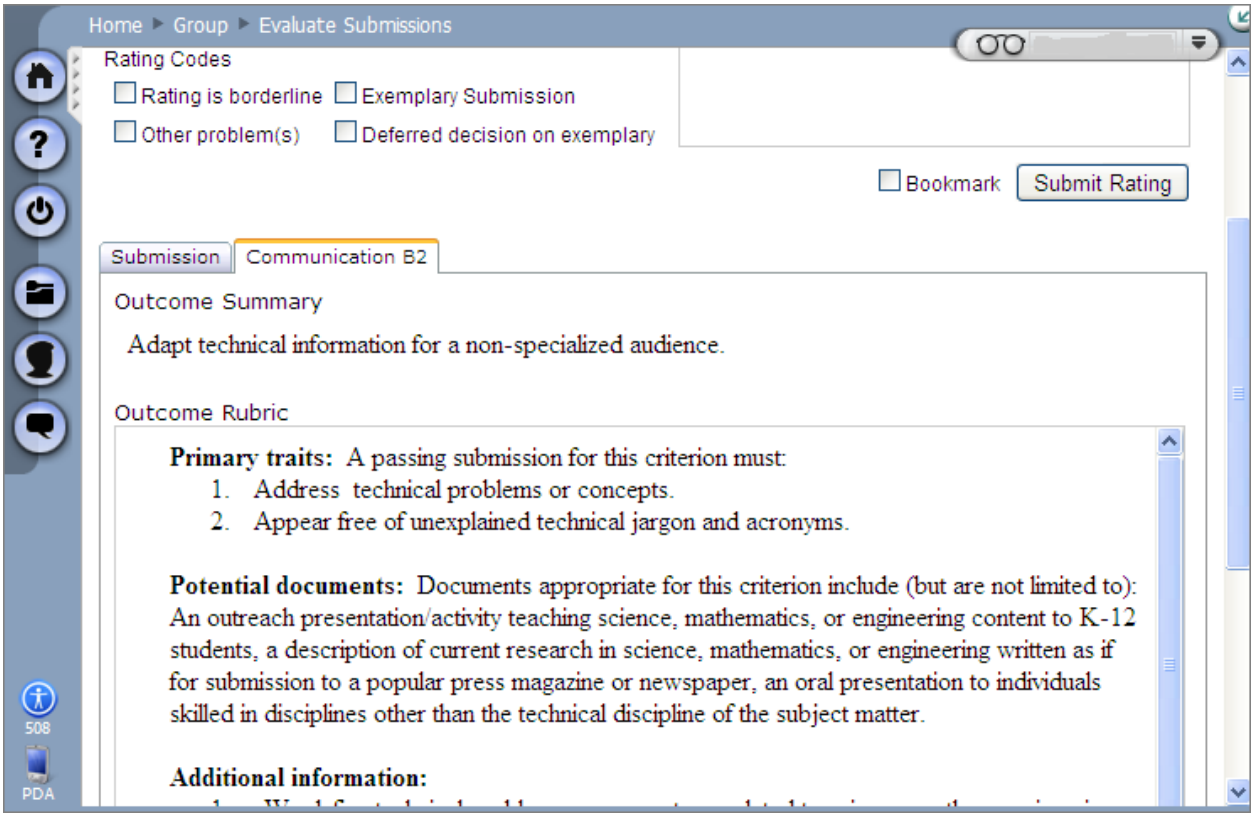
Additional information: None.

APPENDIX C: ROSEVALUATION TOOL SCREENSHOTS

Screenshots of the RosEvaluation Tool

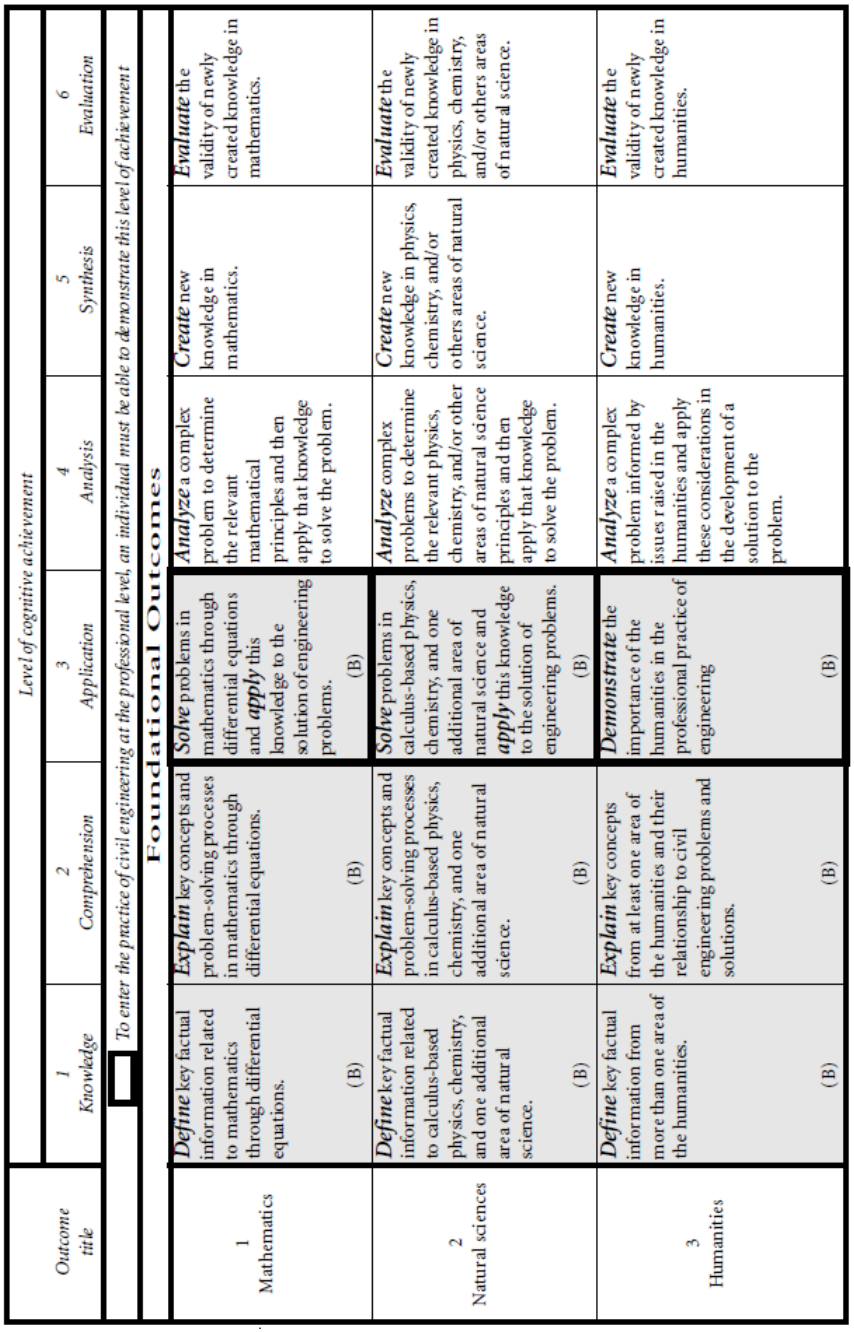



Home ▸ Group ▸ Evaluate Submissions
Rate Submissions
Document #: 4 Communication B2
75 rated of 75 in session - 45 rated by you
Previous
Next
Session List
Rater Comments
Rating Pass
Rating Codes
Rating is borderline
Exemplary Submission
Other problem(s)
Deferred decision on exemplary
Bookmark
Submit Rating
Submission
Communication B2
Submission Title:
Rose Connection
Course Title:
RH330-06 Tech & Professional Communictn
Dropbox Name:
LEARNING OUTCOMES DROPBOX 2
Tech_Description.docx
Student Comments
Here is our Technology Description
Instructor Comments
508
PDA
Internet
100%


Home ▸ Group ▸ Evaluate Submissions
Rating Codes
Rating is borderline
Exemplary Submission
Other problem(s)
Deferred decision on exemplary
Bookmark
Submit Rating
Submission
Communication B2
Outcome Summary
Adapt technical information for a non-specialized audience.
Outcome Rubric
Primary traits: A passing submission for this criterion must:
1. Address technical problems or concepts.
2. Appear free of unexplained technical jargon and acronyms.
Potential documents: Documents appropriate for this criterion include (but are not limited to): An outreach presentation/activity teaching science, mathematics, or engineering content to K-12 students, a description of current research in science, mathematics, or engineering written as if for submission to a popular press magazine or newspaper, an oral presentation to individuals skilled in disciplines other than the technical discipline of the subject matter.
Additional information:
508
PDA

APPENDIX D: ASCE CIVIL ENGINEERING B.O.K.

ASCE Civil Engineering Body of Knowledge for the 21st Century



Outcome title	Level of cognitive achievement					
	1 Knowledge	2 Comprehension	3 Application	4 Analysis	5 Synthesis	6 Evaluation
	☐ To enter the practice of civil engineering at the professional level, an individual must be able to demonstrate this level of achievement					
Foundational Outcomes						
1 Mathematics	***Define*** key factual information related to mathematics through differential equations. (B)	***Explain*** key concepts and problem-solving processes in mathematics through differential equations. (B)	***Solve*** problems in mathematics through differential equations and ***apply*** this knowledge to the solution of engineering problems. (B)	***Analyze*** a complex problem to determine the relevant mathematical principles and then apply that knowledge to solve the problem.	***Create*** new knowledge in mathematics.	***Evaluate*** the validity of newly created knowledge in mathematics.
2 Natural sciences	***Define*** key factual information related to calculus-based physics, chemistry, and one additional area of natural science. (B)	***Explain*** key concepts and problem-solving processes in calculus-based physics, chemistry, and one additional area of natural science. (B)	***Solve*** problems in calculus-based physics, chemistry, and one additional area of natural science and ***apply*** this knowledge to the solution of engineering problems. (B)	***Analyze*** complex problems to determine the relevant physics, chemistry, and/or other areas of natural science principles and then apply that knowledge to solve the problem.	***Create*** new knowledge in physics, chemistry, and/or others areas of natural science.	***Evaluate*** the validity of newly created knowledge in physics, chemistry, and/or others areas of natural science.
3 Humanities	***Define*** key factual information from more than one area of the humanities. (B)	***Explain*** key concepts from at least one area of the humanities and their relationship to civil engineering problems and solutions. (B)	***Demonstrate*** the importance of the humanities in the professional practice of engineering (B)	***Analyze*** a complex problem informed by issues raised in the humanities and apply these considerations in the development of a solution to the problem.	***Create*** new knowledge in humanities.	***Evaluate*** the validity of newly created knowledge in humanities.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
4 Social sciences	***Define*** key factual information from more than one area of social sciences. (B)	***Explain*** key concepts from at least one area of the social sciences and their relationship to civil engineering problems and solutions. (B)	***Demonstrate*** the incorporation of social sciences knowledge into the professional practice of engineering. (B)	***Analyze*** a complex problem incorporating social science knowledge and then apply that knowledge in the development of a solution to the problem.	***Create*** new knowledge in social sciences.	***Evaluate*** the validity of newly created knowledge in social sciences.
Technical Outcomes						
5 Materials science	***Define*** key factual information related to materials science within the context of civil engineering. (B)	***Explain*** key concepts and problem-solving processes in materials science within the context of civil engineering. (B)	***Use*** knowledge of materials science to ***solve*** problems appropriate to civil engineering. (B)	***Analyze*** a complex problem to determine the relevant materials science principles, and then apply that knowledge to solve the problem.	***Create*** new knowledge in materials science.	***Evaluate*** the validity of newly created knowledge in materials science.
6 Mechanics	***Define*** key factual information related to solid and fluid mechanics. (B)	***Explain*** key concepts and problem-solving processes in solid and fluid mechanics. (B)	***Solve*** problems in solid and fluid mechanics. (B)	***Analyze*** and solve problems in solid and fluid mechanics. (B)	***Create*** new knowledge in mechanics.	***Evaluate*** the validity of newly created knowledge in mechanics.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
7 Experiments	***Identify*** the procedures and equipment necessary to conduct civil engineering experiments in more than one of the technical areas of civil engineering. (B)	***Explain*** the purpose, procedures, equipment, and practical applications of experiments spanning more than one of the technical areas of civil engineering. (B)	***Conduct*** experiments in one or across more than one of the technical areas of civil engineering according to established procedures and ***report*** the results. (B)	***Analyze*** the results of experiments and evaluate the accuracy of the results within the known boundaries of the tests and materials in or across more than one of the technical areas of civil engineering. (B)	*Specify* an experiment to meet a need, conduct the experiment, and analyze and *explain* the resulting data. (M/30)	***Evaluate*** the effectiveness of a designed experiment in meeting an ill-defined real-world need.
8 Problem recognition and solving	***Identify*** key factual information related to engineering problem recognition, problem solving, and applicable engineering techniques and tools. (B)	***Explain*** key concepts related to problem recognition, problem articulation, and problem-solving processes, and how engineering techniques and tools are applied to solve problems. (B)	***Develop*** problem statements and ***solve*** well-defined fundamental civil engineering problems by ***applying*** appropriate techniques and tools. (B)	***Formulate*** and solve an ill-defined engineering problem appropriate to civil engineering by ***selecting*** and applying appropriate techniques and tools. (M/30)	***Synthesize*** the solution to an ill-defined engineering problem into a broader context that may include public policy, social impact, or business objectives.	***Compare*** the initial and final problem statements, the effectiveness of alternative techniques and tools, and ***evaluate*** the effectiveness of the solution.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
9 Design	***Define*** engineering design; ***list*** the major steps in the engineering design process; and ***list*** constraints that affect the process and products of engineering design. (B)	***Describe*** the engineering design process; ***explain*** how real-world constraints affect the process and products of engineering design. (B)	***Apply*** the design process to meet a well-defined set of requirements and constraints. (B)	***Analyze*** a system or process to determine requirements and constraints. (B)	***Design*** a system or process to meet desired needs within such realistic constraints as economic, environmental, social, political, ethical, health and safety, constructability, and sustainability. (B)	***Evaluate*** the design of a complex system, component, or process and ***assess*** compliance with customary standards of practice, user's and project's needs, and relevant constraints. (E)
10 Sustainability	***Define*** key aspects of sustainability relative to engineering phenomena, society at large, and its dependence on natural resources; and relative to the ethical obligation of the professional engineer. (B)	***Explain*** key properties of sustainability, and their scientific bases, as they pertain to engineered works and services. (B)	***Apply*** the principles of sustainability to the design of traditional and emergent engineering systems. (B)	***Analyze*** systems of engineered works, whether traditional or emergent, for sustainable performance. (E)	***Design*** a complex system, process, or project to perform sustainably. ***Develop*** new, more sustainable technology. ***Create*** new knowledge or forms of analysis in areas in which scientific knowledge limits sustainable design.	***Evaluate*** the sustainability of complex systems, whether proposed or existing.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
11 Contemporary issues and historical perspectives	***Identify*** economic, environmental, political, societal, and historical aspects in engineering. (B)	***Describe*** the influence of historical and contemporary issues on the identification, formulation, and solution of engineering problems and ***describe*** the influence of engineering solutions on the economy, environment, political landscape, and society. (B)	Drawing upon a broad education, ***explain*** the impact of historical and contemporary issues on the identification, formulation, and solution of engineering problems and ***explain*** the impact of engineering solutions on the economy, environment, political landscape, and society. (B)	***Analyze*** the impact of historical and contemporary issues on the identification, formulation, and solution of engineering problems and ***analyze*** the impact of engineering solutions on the economy, environment, political landscape, and society. (E)	***Synthesize*** the impacts and relationships among engineering and economic, environmental, political, societal, and historical issues.	***Evaluate*** the impacts and relationships among engineering and historical, contemporary, and emerging issues.
12 Risk and uncertainty	***Recognize*** uncertainties in data and knowledge and ***list*** those relevant to engineering design. (B)	***Distinguish*** between uncertainties that are data-based and those that are knowledge-based and ***explain*** the significance of those uncertainties on the performance and safety of an engineering system. (B)	***Apply*** the principles of probability and statistics to ***solve*** problems containing uncertainties. (B)	***Analyze*** the loading and capacity, and the effects of their respective uncertainties, for a well-defined design and ***illustrate*** the underlying probability of failure (or nonperformance) for a specified failure mode. (E)	***Develop*** criteria (such as required safety factors) for the ill-defined design of an engineered system within an acceptable risk measure.	***Appraise*** a multicomponent system and ***evaluate*** its quantitative risk measure, taking into account the occurrence probability of an adverse event and its potential consequences caused by failure.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
13 Project management	***List*** key management principles. (B)	***Explain*** what a project is and the key aspects of project management. (B)	***Develop*** solutions to well-defined project management problems. (B)	***Formulate*** documents to be incorporated into the project plan. (E)	***Create*** project plans.	***Evaluate*** the effectiveness of a project plan.
14 Breadth in civil engineering areas	***Define*** key factual information related to at least four technical areas appropriate to civil engineering. (B)	***Explain*** key concepts and problem-solving processes in at least four technical areas appropriate to civil engineering. (B)	***Solve*** problems in or across at least four technical areas appropriate to civil engineering. (B)	***Analyze*** and solve well-defined engineering problems in at least four technical areas appropriate to civil engineering. (B)	***Create*** new knowledge that spans more than one technical area appropriate to civil engineering.	***Evaluate*** the validity of newly created knowledge that spans more than one technical area appropriate to civil engineering.
15 Technical specialization	***Define*** key aspects of advanced technical specialization appropriate to civil engineering. (B)	***Explain*** key concepts and problem-solving processes in a traditional or emerging specialized technical area appropriate to civil engineering. (M/30)	***Apply*** specialized tools, technology, or technologies to ***solve*** simple problems in a traditional or emerging specialized technical area of civil engineering. (M/30)	***Analyze*** a complex system or process in a traditional or emerging specialized technical area appropriate to civil engineering. (M/30)	***Design*** a complex system or process or ***create*** new knowledge or technologies in a traditional or emerging advanced specialized technical area appropriate to civil engineering. (M/30)	***Evaluate*** the design of a complex system or process, or ***evaluate*** the validity of newly created knowledge or technologies in a traditional or emerging advanced specialized technical area appropriate to civil engineering. (E)

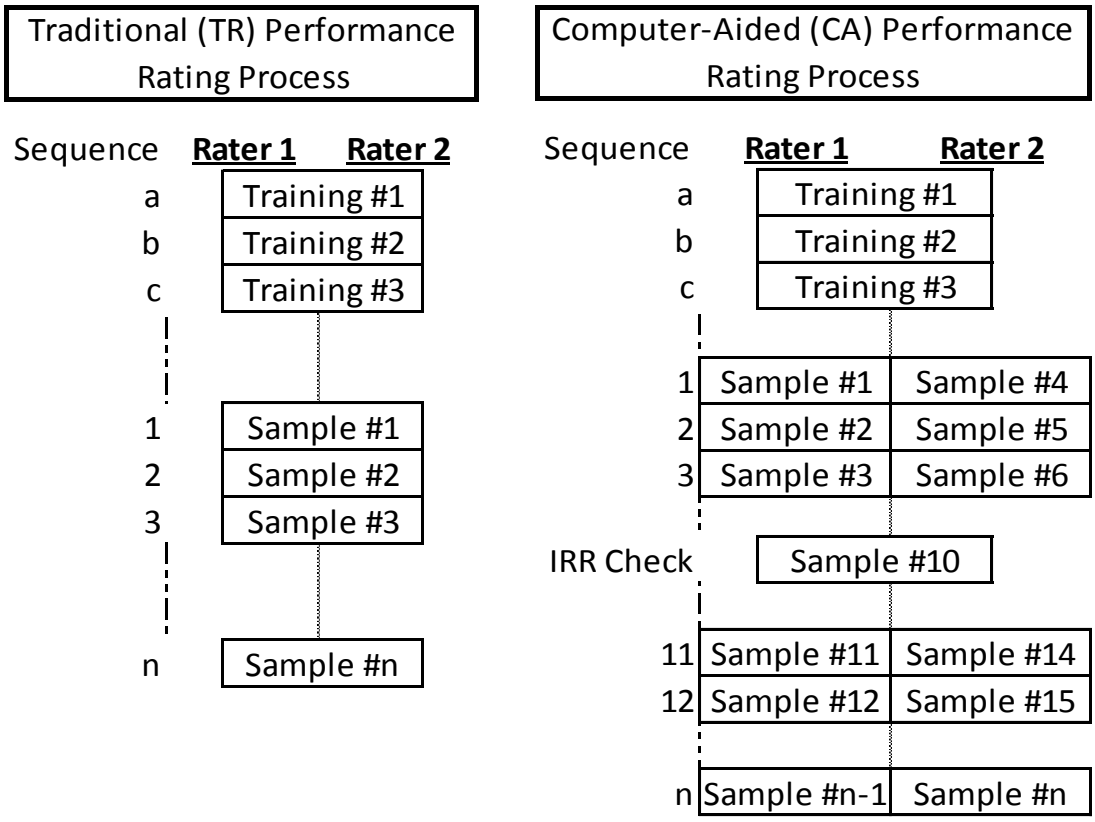
Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
Professional Outcomes						
16 Communication	***List*** the characteristics of effective verbal, written, virtual, and graphical communications. (B)	***Describe*** the characteristics of effective verbal, written, virtual, and graphical communications. (B)	***Apply*** the rules of grammar and composition in verbal and written communications, properly cite sources, and ***use*** appropriate graphical standards in preparing engineering drawings. (B)	***Organize*** and ***deliver*** effective verbal, written, virtual, and graphical communications. (B)	***Plan, compose,*** and ***integrate*** the verbal, written, virtual, and graphical communication of a project to technical and nontechnical audiences. (E)	***Evaluate*** the effectiveness of the integrated verbal, written, virtual, and graphical communication of a project to technical and nontechnical audiences.
17 Public policy	***Describe*** key factual information related to public policy. (B)	***Discuss*** and ***explain*** key concepts and processes involved in public policy. (B)	***Apply*** public policy process techniques to simple public policy problems related to civil engineering works. (E)	***Analyze*** real-world public policy problems on civil engineering projects.	***Develop*** public policy recommendations, and ***create*** or ***adapt*** a system to a real-world situation on civil engineering work programs.	***Evaluate*** the effectiveness of a public policy in a complex, real-world situation associated with large-scale civil engineering initiatives.
18 Business and public administration	***List*** key factual information related to business and public administration. (B)	***Explain*** key concepts and processes used in business and public administration. (B)	***Apply*** business and public administration concepts and processes. (E)	***Analyze*** real-world problems involving business or public administration.	***Create*** or ***adapt*** a system of business or public administration to meet a real-world need.	***Evaluate*** a system of business or public administration in a complex, real-world situation.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
19 Globalization	***Describe*** globalization processes and their impact on professional practice across cultures, languages, or countries. (B)	***Explain*** global issues related to professional practice, infrastructure, environment, and service populations (as they arise across cultures, languages, or countries). (B)	***Organize, formulate,*** and ***solve*** engineering problems within a global context. (B)	***Analyze*** engineering works and services in order to function at a basic level in a global context. (E)	***Develop*** criteria and guidelines to address global issues.	***Evaluate*** different criteria and guidelines in addressing global issues.
20 Leadership	***Define*** leadership and the role of a leader; ***list*** leadership principles and attitudes. (B)	***Explain*** the role of a leader and leadership principles and attitudes. (B)	***Apply*** leadership principles to direct the efforts of a small, homogenous group. (B)	***Organize*** and ***direct*** the efforts of a group. (E)	***Create*** a new organization to accomplish a complex task.	***Evaluate*** the leadership of an organization.
21 Teamwork	***Define*** and ***list*** the key characteristics of effective intradisciplinary and multidisciplinary teams. (B)	***Explain*** the factors affecting the ability of intradisciplinary and multidisciplinary teams to function effectively. (B)	***Function*** effectively as a member of an intradisciplinary team. (B)	***Function*** effectively as a member of a multidisciplinary team. (E)	***Organize*** an intradisciplinary or multidisciplinary team.	***Evaluate*** the composition, organization, and performance of an intradisciplinary or multidisciplinary team.

Outcome title	*Level of cognitive achievement*					
	1 Knowledge	*2 Comprehension*	*3 Application*	*4 Analysis*	*5 Synthesis*	*6 Evaluation*
22 Attitudes	***List*** attitudes supportive of the professional practice of civil engineering. (B)	***Explain*** attitudes supportive of the professional practice of civil engineering. (B)	***Demonstrate*** attitudes supportive of the professional practice of civil engineering. (E)	***Analyze*** a complex task to determine which attitudes are most conducive to its effective accomplishment.	***Create*** an organizational structure that maintains/fosters the development of attitudes conducive to task accomplishment.	***Evaluate*** the attitudes of key members of an organization and ***assess*** the effect of their attitudes on task accomplishment.
23 Lifelong learning	***Define*** lifelong learning. (B)	***Explain*** the need for lifelong learning and ***describe*** the skills required of a lifelong learner. (B)	***Demonstrate*** the ability for self-directed learning. (B)	***Identify*** additional knowledge, skills, and attitudes appropriate for professional practice. (E)	***Plan*** and ***execute*** the acquisition of required expertise appropriate for professional practice. (E)	***Self-assess*** learning processes and ***evaluate*** those processes in light of competing and complex real-world alternatives.
24 Professional and ethical responsibility	***List*** the professional and ethical responsibilities of a civil engineer. (B)	***Explain*** the professional and ethical responsibilities of a civil engineer. (B)	***Apply*** standards of professional and ethical responsibility to determine an appropriate course of action. (B)	***Analyze*** a situation involving multiple conflicting professional and ethical interests to determine an appropriate course of action. (B)	***Synthesize*** studies and experiences to foster professional and ethical conduct. (E)	***Justify*** a solution to an engineering problem based on professional and ethical standards and ***assess*** personal professional and ethical development. (E)

APPENDIX E: PERFORMANCE RATING PROCESS COMPARISONS

Performance Rating Process Diagrams



APPENDIX F: RATER AGREEMENT INDICES FORMULAE

Table 46 Distribution of Subjects by Rater and Response Category (1, 2)

Rater B	**Rater A**		
	1	**2**	**Total**
1	A	B	**B**1 = A + B
2	C	D	**B**2 = C + D
Total	**A**1 = A + C	**A**2 = B + D	**N**

I. Kappa-like Indices (Gwet, 2008a)

Scott's PI (PI):

———— , where p = (A + D) / N and ———— ————

Cohen's Kappa (KAPPA):

———— , where p = (A + D) / N and — — — —

G-Index (GI):

———— , where p = (A + D) / N and q = number of response categories

Gwet's Agreement Coefficient 1 (AC_1):

AC_1 ———— , where p = (A + D) / N and $e(\gamma) = 2P_1(1-P_1)$, where ————

II. Raw Agreement Indices (Uebersax, 2000)

Proportion of Overall Agreement (p):

$$p = (A + D) / N$$

Positive Agreement (pa):

$$pa = 2A / (2A + B + C)$$

Negative Agreement (na):

$$na = 2D / (2D + B + C)$$

Table 47 Distribution of Subjects by Rater and Response Category (+, -)

Rater B	**Rater A**		
	+	-	**Total**
+	A	B	**B**1 = A + B
-	C	D	**B**2 = C + D
Total	**A**1 = A + C	**A**2 = B + D	**N**