ANALYZING THE TECHNICAL QUALITY OF A RUBRIC USED TO ASSESS SCIENCE FAIR PROJECTS

by

MELISSA C. POTTER

A DISSERTATION

Presented to the Department of Educational Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Education

June 2009

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

Melissa Potter

Title:

"Analyzing the Technical Quality of a Rubric Used to Assess Science Fair Projects "

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Education degree in the Department of Educational Leadership by:

Gerald Tindal, Chairperson, Educational Leadership
Paul Yovanoff, Member, Educational Leadership
Leanne Ketterlin Geller, Member, Educational Leadership
David Johnson, Outside Member, Chemistry

and Richard Linton, Vice President for Research and Graduate Studies/Dean of the Graduate School for the University of Oregon.

June 13, 2009

Original approval signatures are on file with the Graduate School and the University of Oregon Libraries.

An Abstract of the Dissertation of

Melissa C. Potter for the degree of Doctor of Education

in the Department of Educational Methodology, Policy, and Leadership

to be taken June 2009

Title: ANALYZING THE TECHNICAL QUALITY OF A RUBRIC USED TO ASSESS SCIENCE FAIR PROJECTS

Approved: _____
Dr. Gerald Tindal

Presenting science fair projects gave students an opportunity to complete a performance assessment that comprised a meaningful task focused on process and subject to standards-based assessment. Students presented science inquiry and engineering design projects to judges at a regional science fair. The judges used the domains of the Potter Rubrics to assess the students' work and assigned a Quality score to each project. Using multiple regression, this study found that the mean scores on the Methods and Analysis domains predicted the mean Quality scores. Analyzing the technical quality of the Potter Rubrics addressed some of the measurement and generalizability concerns about performance assessments. Recommendations for future research and implications for practice were examined.

CURRICULUM VITAE

NAME OF AUTHOR: Melissa C. Potter

PLACE OF BIRTH: Portland, Oregon

DATE OF BIRTH: September 2, 1977

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon
Pacific University, Forest Grove, Oregon
Willamette University, Salem, Oregon

DEGREES AWARDED:

Doctor of Education in Educational Leadership, 2009, University of Oregon
Master of Arts in Teaching, 2001, Pacific University
Bachelor of Arts in Biology, 1999, Willamette University

AREAS OF SPECIAL INTEREST:

Science Education
Assessment
Science Staff Development for Elementary School Teachers

PROFESSIONAL EXPERIENCE:

Teacher on Special Assignment, Center for Science Education, Portland State University, Oregon, 2007-2009

Teacher on Special Assignment, Beaverton School District, Oregon, 2007-2009

Science Fair Director, Beaverton School District, Oregon, 2005-2009

Science Teacher, Southridge High School, Beaverton School District, Oregon, 2001-2007

GRANTS, AWARDS AND HONORS:

Math and Science Partnership Grant, Portland State University, Oregon Department of Education, 2007-2009

Connect2SCIENCE Grant, Intel Corporation, 2008

Intel Foundation Grant, Beaverton-Hillsboro Science Expo, 2005-2008

ACKNOWLEDGMENTS

I thank Professors Tindal, Yovanoff, and Ketterlin-Geller for their patience, insight, and expertise in supporting me through the process of finishing this dissertation. Funding for this project came, in part, from the support of the Beaverton School District, Portland State University, and the Intel Corporation. I could not have completed this work without the love and support of my family and friends. My parents provide a foundation of love, hard work, and pride that allow me to believe in myself. Special recognition goes to Nathan, who offers the hope and compassion of a loving partner in life. Special recognition also goes to my son and the dreams he has yet to dream and achieve.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Traditional science fairs represent a model for performance assessments because they encourage students to complete meaningful tasks that focus on the process of solving a problem (Shavelson, Baxter, & Pine, 1991, 1992). Unfortunately, teachers tend to struggle when preparing students to present a project at a science fair because the standards used in the classroom often differ from those used by judges at a science fair.

To alleviate the struggle of preparing students for a science fair, science fairs need to align the performance standards used by teachers to prepare students and the performance standards used by judges to assess students' work. Aligning the standards between what students, teachers, and judges use as criteria for performance will establish a foundation for making performance assessments conducted outside of the classroom more effective at improving learning for students (Haertel, 1999; Messick, 1995). Including rigorous standards for science fairs will also strengthen them as a model of performance assessments.

A common tool used to outline the standards for a performance assessment is a rubric (Flowers, 2006; Hafner & Hafner, 2003; Novak, Herman, & Gearhart, 1996). For this study, the Potter Rubrics combine standards used by teachers in the classroom and

judges at science fairs (International Baccalaureate Organization [IBO], 2001; International Science and Engineering Fair [ISEF], 2007a; Oregon Department of Education [ODE], 2001; Massachusetts Department of Education [MDE], 2006). The Potter Rubrics also hold promise as tools for both summative and formative assessment. The judges will use the Rubrics as a tool for summative assessment on the day of the science fair. The teachers and the students, however, can use the Rubrics as a tool for formative assessment as they prepare and improve the students' projects for the science fair.

For summative assessments, using clear criteria that articulate the knowledge and skills being assessed strengthens a measurement tool like a rubric (Stokking, van der Schaaf, Jaspers, & Erkens, 2004). In addition, using criteria that are consistent and known to the students will help improve the technical adequacy of a rubric. Examining the extent to which patterns of student performance follow predictions can facilitate efforts to determine the technical adequacy of a rubric (Baker, Abedi, Linn, & Niemi, 1995). In this study, the extent to which patterns in student performance, as measured by the Potter Rubrics, predict judges' perception of the quality of the project will be examined.

The purpose of this study is to examine the domains of a rubric that influence judges' decisions when assessing students' work at a science fair. The relationship examined is that between a judge's scores on a rubric for a project and the judge's overall perception of the project's quality. Of particular interest is whether higher scores

in certain domains of the rubric tend to relate to an overall perception of quality over higher scores in other domains. For example, judges might identify projects that scored high in communication as "high quality," whereas high scores in background may not relate to a perception of high quality.

Examining the relationship between judges' scores on a rubric and their overall perception of a project's quality will provide insight into which domains of a science fair project influence judges' perception of quality. Knowing the factors that influence the perception of quality could also help inform science fair staff about possible changes in training judges.

CHAPTER II

LITERATURE REVIEW

Informing Decision-Making for Educators

The process by which we collect data about student performance varies between teachers, schools, and states. Some districts and states depend on students' performance on traditional standardized tests to measure science achievement. Standardized tests also tend to provide stakeholders with the data required for reporting under the No Child Left Behind Act (U.S. Department of Education [DOE], n.d.). Although this type of testing has a long history of extensive psychometrics that supports their use, they often fail to provide teachers with critical information that can help them guide their instruction. As a consequence, a number of alternatives to standardized tests have emerged.

Performance assessments are the most popular alternative for assessing students' knowledge in science. These performance assessments are believed to better reflect students' understanding of complex concepts (Shavelson et al., 1991). Implementing performance assessments on a large scale poses challenges in measurement and generalizability (Haertel, 1999; Shavelson et al., 1992). Nevertheless, performance

assessments used in the classroom can be an ideal alternative to traditional standardized assessments for informing instructional practices of educators.

Performance Assessments

The critical feature of performance assessments is that they focus on students completing concrete, meaningful tasks (Shavelson et al., 1992). Unlike traditional standardized multiple-choice assessments, performance assessments detail expected learning outcomes for students. As a result, performance assessments can promote learning because they detail expected learning outcomes and identify the criteria for success (Wiliam, 2006).

In addition, performance assessments are process-oriented and can be scored to identify the reasonableness of the procedures students use to solve a problem, not just whether or not the students found the right answer. In this manner they provide tasks that balance the needs of assessment along with the needs of teaching. In other words, a performance assessment "makes a good teaching activity, and a good teaching activity makes a good assessment" (Shavelson et al., 1992, p. 22).

Finally, performance assessments hold potential as instruments of standards-based education because of their potential for supporting teaching and learning (Messick, 1995). Incorporating the characteristics of performance assessments with established performance standards allows students to complete meaningful tasks while demonstrating desired learning outcomes.

These three features of performance assessments (use of meaningful tasks that focus on process and are relevant for standards) make them ideal for use in science, particularly in the context of science fairs. Although performance assessments have been informally adopted in most science fair competitions, their evaluation within this context has not been conducted. This study will use the completion of science fair projects as the performance assessment and focus on the standards for science inquiry and engineering design for the performance standards (see Appendices A and B for thorough descriptions of science inquiry and engineering design constructs).

Science Fairs as Performance Assessments

Science fair projects have the three key features of performance assessments: The projects are (a) meaningful to students, (b) focused on process, and (c) based on standards. For most secondary science fairs, the students select their topics. Students often select topics they find interesting and meaningful. Regardless of the topics selected by students, judges examine the process used by students to answer their research questions. For this study, the standards for student performance align with standards for science inquiry and engineering design.

Science fairs provide a promising model for performance assessments because they incorporate positive characteristics of classroom assessments and reduce the negative characteristics of traditional external assessments (Haertel, 1999). For example, in this study, students know when the science fair will occur and they interact

with the audience who will assess their work—two features not typically seen with external assessments but present in classroom assessments (Haertel, 1999). Two more features present at science fairs but not in traditional external assessments are (a) the teachers' and judges' use of the same scoring method for determining the quality of the students' projects and (b) the judges' feedback on students' individual performance. Similar methods for scoring student work and providing individual feedback further reduce the differences between classroom and external assessments. Using the science fair as an external performance assessment reduces differences in purpose, format, and scoring between classroom and external performance assessments, resulting in a more useful form of assessment.

Using a Rubric for Science Fairs

To reduce the difference between assessments occurring within the classroom and the external assessment of the science fair, teachers and judges need to have a shared understanding of a science project's goals. The development of the Potter Rubrics facilitates this common understanding by clearly outlining the performance goals for science inquiry and engineering design. Before the use of the Potter Rubrics, judges used the judging guidelines from the International Science and Engineering Fair (ISEF) to evaluate projects (ISEF, 2007a). Teachers and students, however, used a variety of guidelines to develop a science project. For example, some teachers used the Oregon State Standards (ODE, 2001) for science inquiry and some students used the

guidelines for the ISEF student handbook (ISEF, 2007b). In addition, the ISEF (2007b) guidelines in the student handbook differ from the ISEF (2007a) guidelines for judges. The Potter Rubrics establish common guidelines for judges, teachers, and students to evaluate and develop science projects.

Using a rubric—e.g., the Potter Rubrics—facilitates both formative and summative assessment. For judges, scoring students using the rubric will provide a summative assessment to determine the quality of the students' projects on the day of the fair. For the students and the teachers, however, the rubric scores could provide a formative and a summative assessment tool. For example, many students participating in the science fair are enrolled in a Science Research class. Science Research teachers can use the performance standards outlined in the rubric to facilitate formative assessment while they instruct their students. Teachers could also use the rubric as a summative tool by using their students' scores to improve how they coach next year's students.

In addition, "innovations that include strengthening the practice of formative assessment produce significant and often substantial learning gains" (Black & Wiliam, 1998, p. 140). To increase their performance, students need to (a) understand the goals for student performance at the science fair, (b) have evidence about their present position relative to the goal, and (c) understand how to move closer to the goal (Black & Wiliam, 1998). The Potter Rubrics establish a foundation for communicating the goals

of the science fair, determining the students' absolute level of performance, and providing an opportunity for students to move closer to the goal.

In a study using survey research and analysis of materials submitted by secondary school natural and social science teachers, Stokking et al. (2004) found a lack of clarity in assessment criteria, assignments, and the validity of teachers' assessment of students' research skills. They infer that, for a summative assessment, providing detailed descriptions of the criteria fosters objectivity and using more than one assessor helps control for reliability. The Potter Science Rubric promotes objectivity by providing judges with clear standards for assessing students' science inquiry skills, comparable to the natural science research skills examined by Stokking et al. Additionally, three judges will assess each student project to increase the reliability of the feedback given to students. In summative assessments, like the use of the Potter Rubrics by judges, reliability and validity, along with objectivity and equality, need careful consideration (Stokking et al., 2004).

Domains of a Science Fair Rubric

The science fair examined in this study takes place every year in Oregon. Teachers who bring students to the science fair are familiar with Oregon's standards for science inquiry. Oregon's standards for science inquiry at the secondary level, however, do not require students' level of performance to increase substantially between middle school and high school (Gross et al., 2005). The overall process of science inquiry as

outlined by the Oregon State Standards provides a good starting point, but more rigorous standards for student performance are needed for the secondary level (Gross et al., 2005).

Both districts that participate in the science fair have International Baccalaureate (IB) schools. Therefore, some of the teachers use IB standards in their classrooms. In addition, IB requires internal assessments for science that require students to conduct their own science inquiry project (IBO, 2001). The outline of the IB standards for science inquiry aligns with the Oregon State Standards. The IB standards, however, provide a level of rigor that is missing from the Oregon State Standards.

For this study, students presented science projects at an ISEF-affiliated fair. According to ISEF (2007b), the elements of a successful project include (a) a project data book, (b) a research paper, (c) an abstract, (d) a visual display, and (e) judging. ISEF's criteria for a successful project do not align with accepted standards for science (ISEF, 2007b; National Center for Education Statistics [NCES], 2006; ODE, 2001). The domains of the Potter Science Rubric (Background, Methods, Data Collection, Analysis, and Communication) combine the criteria of the standards used by teachers and the unique aspects of presenting a science project at an ISEF-affiliated fair.

Unlike the Potter Science Rubric, the Potter Engineering Rubric did not have corresponding Oregon state standards or IB standards as examples to use during the development of the rubric. The Massachusetts Department of Education, however, does include engineering design standards in its Science and Technology/Engineering

Curriculum Framework (MDE, 2006; Sneider & Brenninkmeyer, 2006). In addition, the Fordham Institute gave Massachusetts an "A" for their Science and Technology/Engineering Curriculum Framework (Gross et al., 2005). Specifically, Massachusetts received a 3, the highest possible score, for the science inquiry standards, including engineering design. Recently, the Oregon Department of Education, posted science academic content standards, including standards for engineering design, for public review (ODE, 2008). If the final form of the science standards for Oregon includes engineering design standards, then the standards in the Potter Engineering Rubric may need to be reexamined to ensure that teachers can use the rubric in their classrooms.

Assessing Student Projects at a Science Fair

Judges could use a variety of methods for measuring student performance at a science fair. For example, judges could assign an unconstrained amount of points to each project or assign a certain percentage of points given a maximum score. In addition, judges could use either a generic rubric or a topic-specific rubric. The dependability of scores obtained differs for each method used (unconstrained amount of points, percentage of points, generic rubric, or topic-specific rubric). Using generalizability (G) studies and alternative decision (D) studies, Marzano (2002) compared the four methods of classroom assessments for an eighth-grade science test and found the most dependable scores come from using a topic-specific rubric. The

Potter Science Rubric developed for this study provides judges with a topic-specific rubric for science inquiry. Because the Potter Science Rubric focuses on the topic of science inquiry, it provides a foundation for assessing all of the students' projects, regardless of the project's category (e.g., biology, chemistry, or physics).

Determining the Technical Quality of a Rubric

An assessment tool used to judge students' work needs to meet certain quality criteria. Stokking et al. (2004) summarize the quality criteria into four categories: acceptability, practical utility, reliability, and validity. For this study, the acceptability of the assessment relates to the impartial and unambiguous use of a rubric by judges. Practical utility refers to the feasibility of judges using the rubric during the science fair and the ability of the assessment to discern between different levels of performance.

Although the concepts of acceptability and practical utility were considered during the development of the Potter Rubrics, this study focuses primarily on the quality criteria of reliability and validity. According to the summary of Stokking et al. (2004), reliability depends on the consistent use of the assessment between tasks and between raters. Stokking et al. limit the definition of validity to the degree to which an assessment measures what the educator wants to measure. Although this definition does not include the concept of social consequences or making decisions based on an assessment, this definition does include many facets of validity relevant for this study. Of particular interest to this study is the idea that the results from an assessment should

predict the results on other external criteria, or other expected outcomes. An important aspect of this study's analysis of criterion-related evidence is the assumption that both the assessment tool and the external criteria measure the same constructs of science inquiry or engineering design (Crocker & Algina, 2006).

Many approaches can be used to determine the technical quality of a rubric. For this study, patterns of student performance, determined by judges' scores on the rubric, were compared to expected outcomes, the judges' perception of quality. This comparison helps determine the extent to which patterns of student performance, as measured by students' different domain scores on a new rubric, align with each judge's single quality score for each project.

This method for examining the technical adequacy of a rubric resembles the method used by Baker et al. (1995) for a rubric used in a secondary history classroom. Using confirmatory factor analysis, they found moderate correlations between patterns in student performance and predicted performance. Unlike the rubric used in their study, the Potter Rubrics do not designate some domains as "expert" and other domains as "novice." Instead, the domains for the Potter Rubrics establish a foundation for all levels of performance and rely on the different scores in each domain to distinguish between novice and expert performance. In addition, Baker et al. deduced that a lower level of content knowledge among raters might cause a reduction in the technical adequacy of a rubric. Therefore, judges use the Potter Rubrics to assess projects in the

content area of their expertise. For example, judges with physics backgrounds judge physics projects.

Unlike some studies that investigate the technical adequacy of rubrics (Baker et al., 1995; Novak et al., 1996), this study will not focus exclusively on using trained educators as raters. Instead, this study will investigate the strength of a rubric when used by scientists in the field while assessing student work. Similar to the purpose of a study conducted by Hafner and Hafner (2003), who employed college biology students to assess their peers by using a rubric, the intent of this study is to evaluate the effectiveness of a rubric when used by scientists in the setting of a science fair.

To permit the comparison of students' scores, judges assess the different projects by using the same scoring criteria, or rubric. Because the judges will use the Potter Rubrics as a summative tool for determining the quality of students' projects on the day of the fair, the rubrics will be construct-driven instead of task-driven. Using the construct-driven approach leads to more power in making statistical inferences about the students' potential for demonstrating proficiency in other tasks and therefore improves generalizability (Haertel, 1999). In the future, however, teachers may use the rubric as a formative assessment tool, in which case the rubric could be task-driven.

The internal structure of the Potter Rubrics needs to align with the internal structure of the construct domain (Messick, 1995). For this study, the construct domains for the rubrics are science inquiry for the Potter Science Rubric and engineering design for the Potter Engineering Rubric. Determining the consistency between the structure of

the Potter Rubrics and the construct domains of science inquiry and engineering design will provide evidence for using students' scores on the Potter Rubrics to compare the performance of different students, thus contributing to the valid use of the rubrics.

Multiple Linear Regression as a Foundation for Analysis

Once the judges assessed the five domains of each project, the scores were used to examine interrater reliability and the correlation between rubric scores and the judges' perception of quality. Interrater reliability was determined using intraclass correlation coefficients (ICC; McGraw & Wong, 1996). Multiple linear regression was used to examine the relationship between judges' scores on different domains of a rubric and the judges' overall perception of the quality of each project (high, medium, low). The dependent variable is the perception of quality (high, medium, low), and the independent variable is the judges' score for each domain for each project.

Although some studies have used confirmatory factor analysis to examine the technical adequacy of a rubric (Baker et al., 1995; Flowers, 2006), this study uses multiple linear regression analysis to address the influence each domain of the rubric has on judges' perception of quality. Knowing which domain scores better predict judges' assessment of quality provides insight into the future planning of training for judges and the instructional practices of teachers.

CHAPTER III

METHODOLOGY

Setting

All of the data for this study were collected on the day of a district science fair. The district science fair is part of the Intel International Science and Engineering Fair (ISEF) system and the Northwest Science Expo (NWSE), the statewide fair in Oregon. Students at this regional fair with the top three projects in their category (chemistry, biology, plant science, etc.) continue on to present their projects at the state science fair, and six projects are selected to compete at ISEF.

Approximately 180 students presented 140 projects on the day of the science fair. Some of the projects were team projects with two or three students presenting the same project.

The students who presented their science and engineering projects on the day of the science fair attended school in one of two school districts. School District A is larger (38,000 students) than School District B (20,000 students); however, in terms of students, community, and teachers, both districts have similar demographics.

Participants: Qualification of Judges

The judges volunteered to spend a day assessing students' projects. All of the judges were adults who work in a variety of professions. The fair system required judges to have an M.A., M.S., or Ph.D. in the category they were selected to judge, or in a closely related field.

On the day of the science fair, each judge was assigned to a category (i.e., Engineering: Mechanical or Plant Sciences) based on their preference or experience as indicated by an online judges' registration system. Throughout the day, each judge assessed between 5 and 15 projects, depending on how many student projects were entered in each category.

Research Design

This study was based on a correlational research design. The relationship examined was that between students' scores on the Potter Rubrics and the judges' perception of quality (high, medium, low). The dependent variable was the perception of quality (high, medium, low), and the independent variable was the score each judge gave for each project in each domain.

Procedures: Schedule for Judges

In the morning before the judges started to assess students' projects, they attended a judges' orientation session. During the orientation session, they were given

the option to use the Potter Rubrics as a tool for assessing the students' projects. A rationale for using the rubric was presented to all of the judges along with training on how to use the Potter Rubrics. When judges volunteered to use the rubrics to assess students' work they also volunteered to participate in this study.

After the judges' orientation session, the judges in each category met to determine which projects each judge would assess. In categories with few projects, each judge assessed all of the projects; in larger categories the judges divided up the projects so that each judge assessed 10 to 15 projects. At a minimum, three different judges assessed each project. Once they had their assigned list of projects, the judges assessed the students' projects without the students' presence. During the assessment time without the students, the judges examined the students' display board, paperwork, and research plan. The judges who opted to use the Potter Rubrics began considering how each project would score on the rubrics. Judges assessing science inquiry projects used the Potter Science Rubric, and the judges assessing engineering design projects used the Potter Engineering Rubric.

After the judges spent 90 minutes reviewing the projects without the students, the students entered the exhibit hall and stood next to their posters. The judges then rotated to different projects on their assigned list and interviewed each student. While the judges interviewed students, those who used the Potter Rubrics considered the criteria contained in the rubrics and used them to assess the students' projects.

On each of the rubrics used, the judges recorded their judge number (e.g., PSJ 2 for Plant Sciences Judge 2) and the students' project identification number (e.g., PS 032 for Plant Sciences Project 32). The judges put their judge numbers (e.g., PSJ 2) on the rubrics rather than their names. There was no method for determining which judge was assigned to which number. At no point did two judges interview the same student at the same time. After 2 hours of interviewing students, the judges met for lunch with the other judges in their category. During lunch, the judges in each category discussed the students' projects and completed a Judges' Survey. The judges also confirmed that three judges had seen each project; if any projects had been reviewed less than three times, the judges followed lunch with reviews of those projects.

For 2 hours after lunch, the judges continued to visit different projects and interview students. After the 2-hour afternoon session, the judges in each category met to determine which projects earned first, second, and third places in the category. The judges did not need to use the scores on the rubrics to determine first, second, and third places. The judges also assigned a quality score (high, medium, low) to each project they reviewed during the day. At the end of the day, all of the rubrics for the projects, the rank ordering of the projects, the Judges' Survey, and the quality scores were collected.

Measures

Description of Rubric

The judges used different rubrics to assess science projects and engineering projects. Although the two rubrics have some of the same domains, the purpose and process of a science project differs considerably from the purpose and process of an engineering project. The different rubrics for science and engineering projects reflect the differences in purpose and process for the two types of projects.

Both the science and engineering rubrics have five domains. Under each domain is a list of "what to look for," or criteria to observe when assessing that domain. For example, under the domain "Background," the criteria were (a) description of background, (b) explanation of research question(s), and (c) explanation of the purpose of the project (see Appendix A). The judges gave each student in each domain a score of 1-3, with 1 indicating poor performance and 3 indicating high-quality performance. Each criterion had three different descriptions for the three different levels of performance.

For the science projects, the domains of the rubric were (a) Background, (b) Methods, (c) Data Collection, (d) Analysis, and (e) Communication. The first four domains mirror the domains of the Oregon Department of Education's rubric for assessing science inquiry. The lists of "what to look for" under each domain of the Potter Science Rubric mirror the more rigorous standards established by the

International Baccalaureate program. The last domain of Communication incorporates many of the criteria established by the International Science and Engineering Fair system for communication. For the engineering rubric, the domains of the rubric were (a) Definition of the Problem, (b) Design of the Solution, (c) Data Collection, (d) Analysis, and (e) Communication (see Appendix B). The engineering domains mirror those established in the Potter Science Rubric and incorporate the criteria established by the Massachusetts Department of Education for engineering design.

Table 1 outlines the domains for both rubrics. The two domains that focus on the purpose and process of the projects differ. However, the three domains that focus on data collection, analysis, and communication are the same. The criteria under the domains of data collection and analysis, however, are different between the two rubrics. Both types of projects collect and analyze data, but the types of data collected and the purpose of the analysis differ between a science project and an engineering project. Therefore, the domains have the same title but different criteria. The domain "Communication" and the corresponding criteria for this domain are the same for both rubrics.

Development of Rubric

Four resources were used to develop the first draft of the Potter Rubrics: (a) the Oregon State Standards for science inquiry, (b) the International Baccalaureate standards for science internal assessments, (c) the International Science and Engineering

TABLE 1. Domains for Science and Engineering Rubrics

Science rubric	Engineering rubric
Background	Define the problem
Methods	Design of solution
Data collection	Data collection
Analyzing	Analyzing
Communication	Communication

Fair (ISEF) judging guidelines, and (d) the Massachusetts Science and Technology Framework. A draft of both the Potter Science Rubrics and the Potter Engineering Rubrics were presented to two panels of judges to collect validity evidences. The rubrics were revised after each review session. Each judge on the panel had a Ph.D. in a science content area, a professional engineering license, or five or more years' experience teaching in a secondary school.

Changes to the rubrics during the review process included reducing the number of performance levels for each domain, maintaining consistent "what to look for" descriptions across all of the performance levels for each domain, and refining the descriptions of the "what to look for" criteria. Originally, the rubrics had five different performance levels to describe each domain. In general, all of the reviewers suggested reducing the number of performance levels. The final rubrics, with three performance levels, received the most support from the reviewers and reflect a similar format as the International Baccalaureate rubrics for internal assessments in science. In addition, the reviewers suggested that the rubrics should describe each "what to look for" criterion at

each performance level, rather than some of those criteria appearing only at one or two performance levels. Finally, the reviewers provided suggestions for phrasing that helped clarify the descriptions in the rubrics and reduced some education jargon.

Data

The data collected for this study came from the judges. The judges' data included the scores given to students on each domain of the rubrics and the quality scores given to each project by each judge. None of the participants' names were associated with the data. Identification numbers were used to distinguish between judges (PSJ 2) and students' projects (PS 032). The data collected came from three groups of judges: (a) engineering, (b) environmental, and (c) cellular molecular. In each of the three groups of judges, the judges assessed projects from multiple categories. For example, in engineering, the judges assessed both material and electrical engineering projects, and in the cellular molecular category, the judges assessed biochemistry, microbiology, and medicine and health projects.

Data Analysis

Interrater Reliability Using Intraclass Correlation

Intraclass Correlation (ICC) was used to determine the interrater reliability for this study. This study used ICC because there were multiple raters, the rating scale included more than two values, and this study assumed that the rating scale consisted of

interval-level data. The data used to determine interrater reliability came from a pilot study conducted in February 2008. Specifically, the data included information collected from five engineering judges who assessed 21 projects.

The results from the ICC suggest that the raters, or judges, were in agreement for four out of five of the domains (Definition of the Problem, Design of the Solution, Analysis, Communication). For the domain of Data Collection, a low average measure ICC (0.47) and a low coefficient alpha (0.44) were obtained. Acceptable alpha is usually at least 0.6.

Correlation Between Rubric Scores and Perception of Quality

Multiple linear regression was used to examine the relationship between judges' scores on different domains of a rubric and the judges' overall perception of the quality of each project (high, medium, low). The dependent variable, or outcome, was the perception of quality (high, medium, low), and the independent variable, or predictor, was the score each judge gave for each project in each domain. Once the data were collected, averaging the scores on each domain for each project consolidated the data. In addition, the quality scores for each project were averaged.

Averaging the scores increased the variability in scores and allowed the analysis of individual projects regardless of the judge who assessed the project. Obtaining reasonable interrater reliability during the pilot program in February 2008 permitted the scores to be averaged.

Descriptive statistics, correlations, and multiple regressions were used to analyze the data. Using the multiple regression equation

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

could predict the quality score (y) by using the students' average scores from the domains of background domain (x_1), methods (x_2), data collection (x_3), analysis (x_4), and communication (x_5).

CHAPTER IV

RESULTS

The data for this study come from the scores that judges gave to students during a science fair. Three groups of judges—engineering, environmental, and cellular/molecular—assessed three groups of projects. The fair system required judges to have an M.A., M.S., or Ph.D. in the category they were selected to judge, or in a closely related field. In general, the judges had more than 5 years of experience in their careers and less than 5 years of experience judging science fairs (see Table 2).

The judges assigned each project five domain scores, one indicating poor performance and three indicating excellent performance. The judges also assigned a Quality score to each project, indicating if they thought the project was low quality (1) or high quality (3). Each project was assessed at least three times by three different judges. The five domain scores and the Quality scores for each project were averaged for each student. As a result, each project had an aggregate score for each of the five domains and an aggregate Quality score.

Interrater Reliability

Intraclass Correlation (ICC) was used to determine the interrater reliability for this study. The results from the ICC suggest that the raters, or judges, were in agreement

TABLE 2. Demographics of Science Fair Judges

	Engineering ($n = 5$)	Environmental ($n = 5$)	Cellular/ molecular ($n = 9$)	Combined ($n = 19$)
Gender				
Male	4	4	4	12
Female	1	1	5	7
Years of judging experience				
0-2	3	3	8	14
3-4	2	2	1	5
5 or more	0	0	0	0
Years of career experience				
0-2	0	0	2	2
3-4	0	0	0	0
5 or more	5	5	7	17
Race[a]				
American Indian or Alaskan Native	0	0	0	0
Asian	2	1	0	3
Black or African American	0	0	0	0
Native American or other Pacific Islander	0	0	0	0
White	3	3	6	12
Some other race	0	0	0	0

Note. The fair system required judges to have an M.A., M.S., or Ph.D. in the category they were selected to judge, or in a closely related field.

[a]Responding to this question on the survey was optional.

in all five of the engineering rubric domains. For each of the domains, the average measure ICC was greater than 0.78 ($p < .001$) and the Cronbach's Alpha was 0.86 or greater. Interrater reliability was also examined in terms of exact agreement and agreement within one score point. Agreement was over 72% between raters within one score point (see Table 3).

TABLE 3. Percent Agreement of Judges

Category	Background	Methods	Data collection	Analysis	Communication
			Percentage exact agreement		
Engineering	9.1	9.1	27.3	27.3	9.1
Cellular	26.1	13.0	17.4	21.7	21.7
Environmental	27.8	33.3	22.2	22.2	22.2
			Percentage exact agreement ±1 score point		
Engineering	81.8	81.8	72.7	72.7	90.9
Cellular	100.0	95.7	91.3	95.7	95.7
Environmental	83.3	100.0	94.4	88.9	94.4

Descriptive Statistics

The engineering judges tended to give each domain a score of less than 2 (see Table 4). The standard deviation for all of the engineering scores is greater than 0.6. In the Background and Analysis categories, judges' scores averaged under 2, but in the Methods, Data Collection, and Communication categories, the scores judges gave averaged above 2. All but one of the standard deviations for the environmental scores is above 0.5. Unlike the engineering judges, the cellular/molecular judges tended to give scores greater than 2 for each domain ($SD > 0.3$).

TABLE 4. Descriptive Statistics for Judges' Assessment of Science Fair Projects

	Engineering (n = 11)	Environmental (n = 18)	Cellular/ molecular (n = 23)	Combined (n = 51)
Background				
Mean	1.95	1.90	2.22	2.06
Standard deviation	.62	.59	.41	.54
Skewness	.30	.09	-.18	-.18
Methods				
Mean	1.85	2.03	2.27	2.10
Standard deviation	.65	.56	.38	.53
Skewness	.35	-.13	-.73	-.42
Data collection				
Mean	1.70	2.06	2.20	2.05
Standard deviation	.74	.57	.36	.55
Skewness	.43	.10	-.51	-.38
Analysis				
Mean	1.64	1.69	2.22	1.91
Standard deviation	.66	.63	.56	.66
Skewness	.57	.71	-.46	.06
Communication				
Mean	2.02	2.24	2.39	2.26
Standard deviation	.69	.49	.43	.52
Skewness	.06	-.13	-.52	-.44
Quality				
Mean	1.73	2.04	2.35	2.12
Standard deviation	.69	.71	.63	.72
Skewness	.54	-.26	-.28	-.22

Note. A score of 1 indicates poor performance and 3 indicates excellent performance.

The three groups of judges also assigned each project a quality score that conveyed their perception of the overall quality of each project. The engineering judges tended to have a lower perception of quality (M = 1.73, SD = 0.69) than either the

environmental judges ($M = 2.04$, $SD = 0.71$) or the cellular/molecular judges ($M = 2.35$, $SD = 0.63$).

The Relationship Between Quality Scores and Domain Scores

To examine the relationship between each of the domains and quality, I ran bivariate regressions. In general, all of the bivariate regression analyses indicate a positive relationship between the students' average domain scores and their average quality scores (see Figures 1, 2, 3, 4, and 5). The bivariate regression analyses for the Methods and Analysis domains explain more than 70% of the variation seen in the students' mean quality scores (see Figures 2 and 4). The bivariate regression analysis for the Methods domain predicts that, on average, for every point students receive for their average Methods score they receive a 1.0 increase in the average Quality score from the judges (see Figure 2). This model also predicts that, on average, a one-point increase in a student's Analysis score results in a 1.1 increase in the Quality score (see Figure 4).

I conducted a multivariate regression analysis to explore the relationship between the scores given on each domain of a rubric and the judges' perception of quality (see Appendix C for zero order correlation matrix). I included as a covariate which group of judges (engineering, environment, or cellular/molecular) scored each project as a covariate.

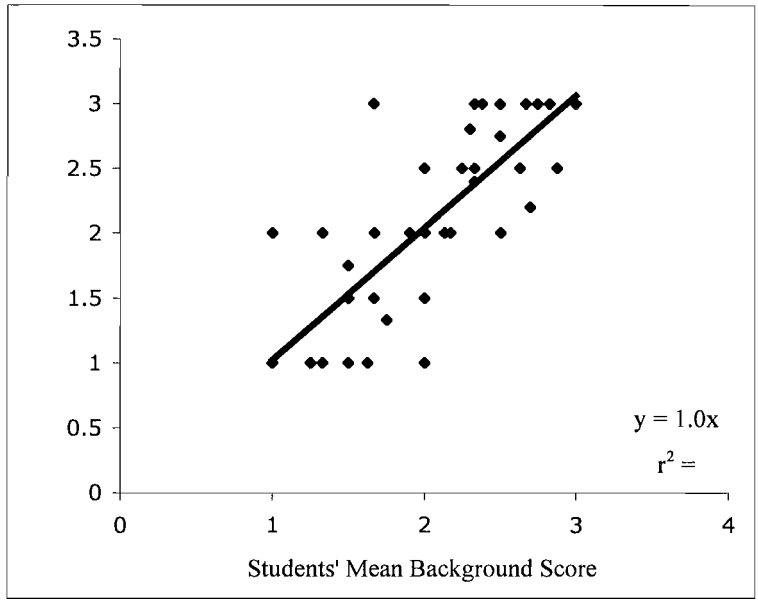

FIGURE 1. Relationship between students' Quality scores and students' Background scores.

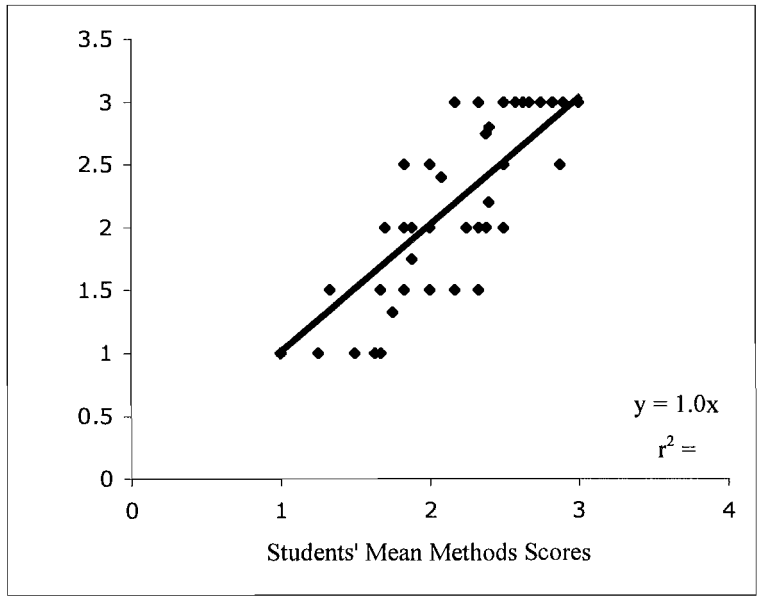

FIGURE 2. Relationship between students' Quality scores and students' Method scores.

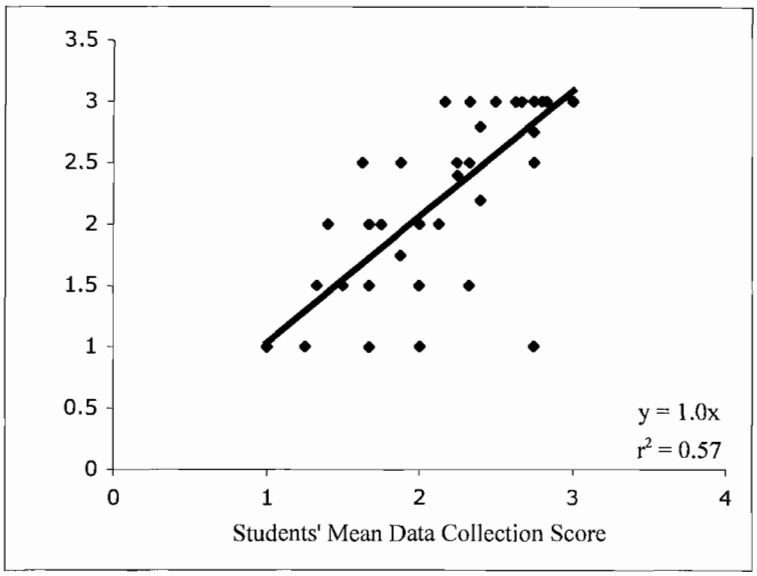

FIGURE 3. Relationship between students' Quality scores and students' Data Collection scores.

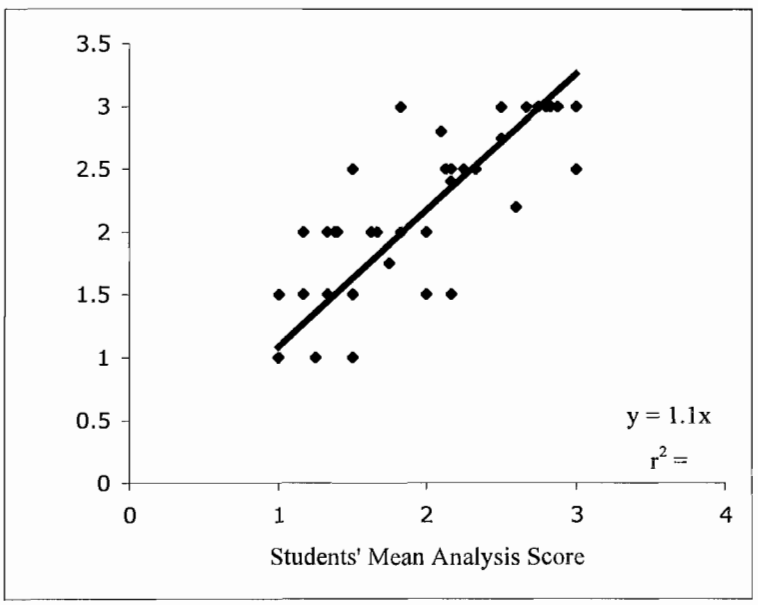

FIGURE 4. Relationship between students' Quality scores and students' Analysis scores.

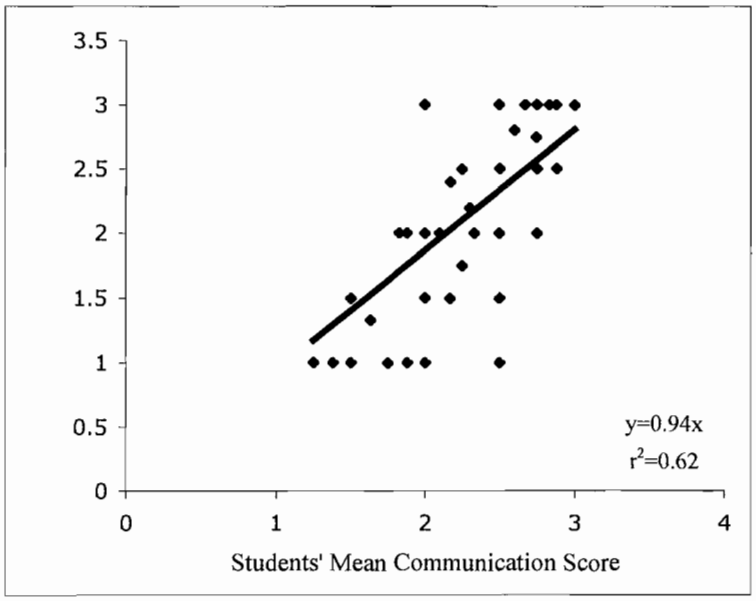

FIGURE 5. Relationship between students' Quality scores and students' Communication scores.

The model had predictive power ($p < .001$). Scores given on the different domains of the Potter Rubric explain approximately 98% of the variation in judges' perception of quality scores ($R^2 = .98$; see Table 5). A relationship exists between students' scores on Methods and the judges' perception of quality ($p < .01$). The findings also suggest a relationship between students' scores on Analysis and the judges' perception of quality ($p < .05$).

To gain a better understanding of the role certain domains play in the prediction of quality scores, I conducted a reduced regression that included the domains that significantly predicted quality scores: Methods, Analysis, and Communication (see Table 5). The significance of the relationship between

TABLE 5. Multiple Linear Regression Summary Predicting Overall Quality With Domains

Variable	*B*	*SE*	Standardized Beta	Sig	Partial Correlation
		Full Multivariate Regression			
Constant	—	—	—	—	—
Background	-.26	.21	-.25	.224	-.03
Method	.59***	.19	.57	.003	.07
Data collection	.00	.17	-.04	.813	-.01
Analysis	.55**	.21	.50	.013	.05
Communication	.37*	.20	.37	.088	.04
In engineering	-.48	.28	-.10	.090	-.04
In environment	-.29	.29	-.08	.332	-.02
In cellular	-.37	.27	-.11	.179	-.03
$R^2 = .98$***, $SE = .33$					
		Reduced Multivariate Regression			
Constant	—	—	—	—	—
Method	.50***	.17	.49	.006	.06
Analysis	.44**	.19	.40	.022	.05
Communication	.27	.18	.29	.144	.03
In engineering	-.57	.27	-.11	.041	-.04
In environment	-.33	.28	-.09	.253	-.02
In cellular	-.42	.27	-.13	.127	-.03
$R^2 = .98$***, $SE = .32$					

*$p < .10$. **$p < .05$. ***$p < .01$.

Communication and Quality ($p \leq .01$) was less than the relationship between either Methods or Analysis and Quality. Nevertheless, I included the Communication in the reduced regression because the small sample size in this study increases the sensitivity of statistical significance. Including Communication in the reduced regression prevents the possibility of making a Type II error. The reduced regression indicates that the students' average scores in the Methods and Analysis domains predict the students' average quality scores ($p < .05$).

CHAPTER V

CONCLUSION

Major Findings

This study focused on the capacity of the judges' rubric scores to predict the overall quality score for each project, one facet of validity. Although not all of the domain scores predicted judges' perception of quality, the students' mean scores in the Methods and Analysis domains did predict the students' mean Quality scores. The domain scores for Background, Data Collection, and Communication did not appear to influence the judges' perception of quality.

Study Limitations

The Potter Rubrics measured the construct of science inquiry and engineering design. During the development stages of the rubrics, two panels of experts examined the rubrics and rated how well the rubrics described each domain of science inquiry and engineering design and how well the rubrics described the different levels of performance students could achieve for each domain. Feedback from these expert panels provided the information used to make revisions to the Potter Rubrics. A reasonable

attempt, therefore, was made to effectively explain the constructs of science inquiry and engineering design for this study.

This study used only one method, or mono-method bias, for collecting data. Due to the design of the study, the judges provided both the domain scores and the quality scores. Using only one method to measure the dependent variable may limit the generalizability of this study's inferences to studies that use a different method for determining quality (Shavelson et al., 1992).

Another method of measuring quality could have been determining if the project advanced to the State Science Fair from the regional science fair used for this study. Using this method may have provided insight into the relationship between the higher domain scores and quality, but it would not have provided an opportunity to examine the relationship between the lower domain scores and quality.

Internal Validity

Threats to internal validity in this correlational design included small sample size, order effect, generalizability, and rater reliability. The raters, or judges, for this study examined approximately 15 projects. The small sample size for each rater made it difficult to establish an estimate of the true level of agreement between raters (Novak et al., 1996). In addition, the order in which the judges viewed the projects may have influenced their ratings. This outcome may have confounded the results by students becoming better as a function of time to prepare or by the raters becoming more

proficient in using the rubrics over time (Hafner & Hafner, 2003). However, unlike the protocol used in the Hafner and Hafner study, all students in this study presented multiple times throughout the day instead of just once, which might have reduced the influence of presentation order on the scores given to students by the judges.

This study examined one performance for each student. Using data collected from numerous student-performance assessments would have increased the equality and generalizability of the performance level for each student (Stokking et al., 2004). To manage the influence of different groups of judges assessing different groups of projects, the multivariate regression analysis included which group of judges (engineering, environment, or cellular/molecular) scored each project as a covariate. The results suggested a relationship between both Methods and Analysis scores and the judges' Quality scores.

Statistical Conclusion Validity

Statistical conclusion validity is limited to the assumptions from multiple regression. Using multiple regression assumes dichotomous or continuous variables. Some might argue that the scale used for this study was categorical. Although the 3-point scale used for both rubric scores and quality scores might be better described as ordinal, this study assumed, for the purpose of analysis, that the scale was intervallic and continuous.

The most significant limitation was the dependency among scores for students across both judges and domains. That is, the same judges rated multiple students and each judge provided multiple ratings (across domains). This design feature, therefore, constrained the variance and violated a basic assumption of independence of error terms. However, the rater reliability results indicated that the judges used the Rubrics consistently. With the assumption that the judges used the Rubrics similarly, the domain scores and Quality scores were averaged for each student, thus increasing the variance and focusing the analysis on the domain scores rather than on rater variability.

External Validity

With the design, the judges selected for the study had the option of using the Potter Rubrics to provide feedback to students. The students selected to participate in the science fair did so as an assignment for an elective science course called Science Research, extra credit for another class, an after-school research club, or an elective summer research program. The selection of participants might limit the generalizability of this study's inferences to settings lacking resources similar to a regional science fair.

Conclusions and Explanations

Rater Reliability

Unlike most studies examining the use of a rubric, this study did not use trained educators to assess students' work. The ICC interrater reliability results of this study

indicated that noneducators, when trained, used the engineering rubric consistently to assess secondary students' work (McGraw & Wong, 1996). Additionally, the agreement was high between raters for all categories within one score point.

Unlike some studies, this study used raters, or judges, whose background was related to the category that they judged (Baker et al., 1995). For example, judges with an engineering background judged engineering projects by using the engineering rubric. Ensuring that the raters who used the rubrics had relevant background in their content areas may have increased the consistency in their use of the rubric.

Although the rater reliability from the pilot study indicated that judges did not consistently use the engineering rubric to measure the Data Collection domain, the results from this study indicated that the Data Collection domain did not influence judges' perception of quality.

Science Fairs as Performance Assessments

This study used data collected at a science fair and thus provided an assessment opportunity different from traditional classroom assessments or standardized assessments. The science fair projects gave students a chance to complete a meaningful task focused on process and subject to standards-based performance assessments (Messick, 1995; Shavelson et al., 1992). Using science fair projects had the potential to provide teachers with information to help them guide their instruction and to provide other stakeholders with accountability information.

Analyzing the technical quality of the Potter Rubrics addressed some of the measurement and generalizability concerns about performance assessments (Haertel, 1999; Messick, 1995; Shavelson et al., 1992). For example, identifying the extent to which scores on a performance assessment correlated with external variables confirmed the value of the scores for an applied purpose (Messick, 1995). This study began to address the meaning of the scores obtained when using the Potter Rubrics by identifying which domain scores predicted the overall Quality score assigned to a project.

Additional examination of the correlation of the Potter Rubrics with other tasks representing the same constructs would have addressed the generalizability of the results between tasks (Messick, 1995). Although students for this study researched different topics, they each completed the same task of preparing a science fair project based on either a science inquiry project or an engineering design. To determine the ability of the Potter Rubrics to assess the broad construct domains of either science inquiry or engineering design, the degree of correlation with other tasks representing the same constructs needed to be examined (Novak et al., 1996).

Two other areas that may have influenced the generalizability of performance assessments were the knowledge level of the raters and the instruction given by teachers as students prepared their presentation for the science fair (Baker et al., 1995). The raters, or judges, in this study had an M.A., M.S., or Ph.D. in the category they judged. This study did not, however, examine how the teachers instructed the students during the preparation for the science fair. Knowledge of the various methods used to prepare

students and the possible correlation to scores would have provided critical contextual information for determining the technical quality of the Rubrics.

Development of Potter Rubrics

This study began to address the characteristics of the Potter Rubrics to determine their technical adequacy. Of particular interest were the quality criteria for assessments outlined by Stokking et al. (2004): reliability, validity, acceptability, and practical utility. Examining the interrater reliability of both the pilot study and this study indicated that judges tended to score students similarly while using the rubrics. This examination of reliability suggested that judges used the rubrics consistently.

During the development of the topic-specific rubric, efforts were made to address many facets of validity (Marzano, 2002). For example, as suggested by the expert panel of judges, teachers, scientists, and engineers, the rubrics attempted to sufficiently cover science inquiry and engineering design content. The expert panel also provided feedback on maximizing the specificity of the rubrics and ensuring that the scoring process mirrored the structure of the students' task (Stokking et al., 2004). Their comments also led to the enhancement of the acceptability and practical utility of the rubrics for the raters.

Relationship Between Domain Scores and Quality Scores

This study focused on collecting criterion-related evidence for validity. The critical assumption for examining the relationship between domain scores and Quality scores was that both scores represent students' performance on the same constructs (Crocker & Algina, 2006). Both sets of scores indicated the students' performance on either science inquiry or engineering design. Analyzing the concurrent evidence of both rubric and Quality scores determined the extent to which the students' results on the rubrics overlapped with the Quality scores. Both the rubric scores and the Quality scores measured the same construct.

An explanation for the relationship between Methods and Analysis scores and Quality scores may have resided in the details of the Potter Rubrics. For example, judges looked at the appropriateness of the research plan, the methods for data collection, and the control of variables when assessing the Methods section of the Potter Science Rubric. When using the Potter Engineering Rubric, however, the judges examined the students' exploration of possible alternatives to answering the engineering need or problem, identification of a solution, and creation and testing of a prototype. The components comprising the overall Methods scores might have affected the judges' perception of quality more than the components of other domains. Similarly, the components of the overall Analysis scores might have influenced the judges' perception of quality more than the components of Background and Data Collection.

There was no relationship between Background and Data Collection scores and the Quality scores, or predicted performance (Baker et al., 1995). Perhaps the judges did not value Background as much as Methods and Analysis, or they may have had insufficient time to adequately address the quality of the Background for each project. Although the judges spent 10 minutes interviewing the students at each project, their emphasis was probably related to the students' work and not necessarily on the background for the projects. Analysis scores explained some of the variation in Quality scores, but the Data Collection scores did not explain any variation in Quality scores. The judges may have valued the inferences students made with their data more than the presentation of the data in graphs and tables.

When judges' categories were not included as covariates for the multiple regression, Communication did predict some of the variation seen in the judges' Quality scores. However, when judges' categories were included as covariates, there was no relationship between mean Communication scores and mean Quality scores. The rationale for using the Potter Rubrics was to provide judges, teachers, and students with a common tool regardless of the category or content. Therefore, this study favors the results that included category as a covariate and indicated that Communication did not influence judges' final perception of quality.

The focus of this study was the predictive nature of the judges' rubric scores, one facet of validity. A method of determining the predictive power of an assessment tool is to ascertain whether the results from the tool relate with external criteria

(Messick, 1995; Shavelson et al., 1992; Stokking et al., 2004). In this study, the tools examined were the Potter Rubrics and the alternate criterion was the judges' Quality scores. Although not all of the domain scores predicted judges' perception of quality, the mean scores on the Methods and Analysis domains did predict the mean Quality scores.

Recommendations and Implications

The initial motivation for this study came from the desire that all students, teachers, and judges would use the same tool, or rubric, to prepare and assess projects for a regional science fair. This study sought to examine the relationship such a tool might have with the judges' evaluation of each project. Knowing that two domains on the rubric influence judges' perception of quality can help students and teachers focus on Methods and Analysis in their preparation for the science fair.

The results from this study do not suggest that students can ignore the importance of the other domains of Background, Data Collection, and Communication. The Oregon State Standards for secondary students continue to emphasize the importance of demonstrating the ability to provide quality background, data collection, and communication skills (ODE, 2008). The results from this study indicate, however, that judges, with limited time to evaluate students' work at a science fair, may focus their energy on assessing the Methods and Analysis sections of the students' projects.

The results from this study also suggest that a discrepancy exists between the science inquiry and engineering design standards established by organizations, including the Oregon Department of Education, and the criteria that explains judges' perception of quality. To reconcile this discrepancy, the science fair could modify the judges' orientation to emphasize the need to focus equally on all of the domains represented in the Potter Rubrics. On the other hand, organizations responsible for determining the quality of students' performance in the areas of science and engineering may want to reconsider the standards they use to describe high-quality performances. Along those lines, the term "quality" warrants further consideration.

This study assumed the term "quality" incorporated all of the domains of the Potter Rubrics. A possible alternative, however, is to define "quality" using only a few of the domains. Perhaps science educators only need to consider students' performance in Methods and Analysis when making decisions about student achievement. On the other hand, science educators might consider why they deem the other domains of Background, Data Collection, and Communication important when deciding the performance levels of students. If the goal of accountability systems in education is to measure the quality of students' knowledge and skills, then it is imperative to carefully examine the definition of "quality."

Further studies examining teachers' use of the rubric in their classrooms to facilitate students' learning would provide an interesting opportunity for research (Black & Wiliam, 1998). In addition, studying how students use the judges' feedback on the

rubrics to improve their science projects would provide a better understanding of how to help students improve their work. Using the rubric could also clarify teachers' assessment criteria of students' research skills (Stokking et al., 2004). Finally, raters (teachers, judges, or students) could apply the Potter Rubrics differently to various types of assessments (Novak et al., 1996). The raters' interpretation and use of the rubrics might affect the widespread use of the Potter Rubrics.

The results from this study provide initial insight into the possibility of using a new assessment tool to prepare and assess students' projects for a science fair. Many secondary school standards and programs, including the Oregon State Standards and the International Baccalaureate program, require students to demonstrate skills similar to those skills demonstrated at a science fair. In addition, the Oregon standards for science now include engineering design (ODE, 2008). This study shows that engineering professionals can consistently use an assessment tool, like the Potter Engineering Rubric, to assess secondary students' work. The Potter Engineering Rubric might provide a foundation for the development of a statewide assessment tool that supports teachers' work with students and produces technically sound results.

APPENDIX A

SCIENCE FAIR RUBRIC FOR JUDGES—SCIENCE

Judge ID:_____Project #: _____

Science Fair Rubric for Judges—Science

Please use the following rubric to provide feedback for each student's project. After examining the display board, looking at the research plan, and talking with the student, provide feedback to the student by providing comments and/or scoring the student in each section (Background, Methods, Data Collection, Analysis, and Communication) using the guidelines in the rubric. Also, feel free to circle comments on the rubric that summarize your comments.

Comments (optional):_____

_____

These scores are for feedback only--they do not represent students' final ranking in their category or in "Best of Fair"

Background: _____ Methods: _____ Data Collection: _____ Analysis: _____ Communication: _____

Lab Section **Background** What to look for: ☐ Description of background information ☐ Explanation of research question(s) ☐ Explanation of purpose of project	**Score**		
	1	2	3
	▪ Background information missing, incomplete or only partly relates to the investigation ▪ Research question missing or cannot be directly answered, tested, or clearly explained ▪ The purpose is missing or unclear	▪ Background relates to investigation and accurately uses some supporting literature ▪ Testable research question ▪ Clear description of purpose	▪ Background information is thorough and clear and extensively uses scientific literature ▪ Question or hypothesis focuses on scientific relationships ▪ Clear and focused purpose with an explanation of how it will contribute to a body of knowledge

Judge ID:_____Project #: _____

Methods*	**Score**		
What to look for:	1	2	3
☐ Detailed plan/methods ☐ Appropriate-ness of research plan ☐ Methods for data collection ☐ Control of variables *Consider how much help the student received	▪ Methods are missing or not clearly explained or understood ▪ Methods are not or somewhat related to research question or equipment used inappropriately ▪ Methods allow for the collection of no or some relevant data ▪ Controlled for no or some variables	▪ Methods are clearly explained ▪ Methods are appropriate to answer research question ▪ Design is practical and gives enough information to answer question ▪ Controlled for as many variables as possible	▪ Methods are clear and precise ▪ Research plan is selected based on scientific principles ▪ Design gives enough of the right kind of data and explains relationship ▪ Design thoroughly controls for variables

Lab Section	Score		
Data Collection:	1	2	3
What to look for: ☐ Tables with labels, titles, and units ☐ Graphs (if appropriate) ☐ Transformation of data (calculations, graphs, etc.) to help explain patterns, trends, and an answer to the question.	▪ Tables are not easily understood—missing titles, labels, and/or units ▪ Graphs are unorganized and/or not relevant to the research question (i.e., bar graph when line graph should have been used) ▪ Displays are somewhat appropriate and complete but do not help make interpretations in patterns	▪ Tables are clear and well labeled ▪ Graphs are clear and help answer the research question ▪ Patterns and trends in the data are identified and relate to research question	▪ Tables are thoroughly labeled and annotated as well as clear ▪ Graphs clearly show relationships between variables (and if applicable, supported by the correct and thorough use of statistics) ▪ Patterns and trends are thoroughly discussed and are supported by scientific principles (and if applicable, supported by the correct and thorough use of statistics)

Judge ID:_____Project #: _____

Analysis What to look for: ☐ Results of the investigation ☐ Use science concepts, models and terminology in conclusion. ☐ Review of investigation for possible errors. ☐ Explanation of future studies or uses for project	Score		
	1	2	3
	▪ Conclusions presented are not supported by the data ▪ Connection between conclusion and scientific knowledge or literature is limited ▪ Errors and limitations dealt with in trivial or illogical manner ▪ Explanation of relevance of findings is minimal and ideas for improvement are brief	▪ Clear explanation of results ▪ Connection between conclusion and scientific knowledge and literature is clear and accurate ▪ Clear understanding of the errors and limitations of the project ▪ Presentation of ideas for improvement and relevance of findings	▪ Clear and thorough explanation of the results ▪ Thorough use of scientific literature and data to support conclusions ▪ Errors and limitations are clearly identified and analyzed thoughtfully ▪ Clear outline of "next steps" and relevance of findings
Communication What to look for: ☐ Display board ☐ Forms and research plan ☐ Verbal discussion			
	1	2	3
	▪ Display board is unorganized and difficult to understand ▪ Science fair forms and/or the research plan are missing or unorganized ▪ Student struggles to explain some parts of the project (nervousness is okay)	▪ Display board is easy to understand ▪ Forms and research plan are clear and easy to follow ▪ Student clearly explains all aspects of the project	▪ Display board clearly explains all aspects of the project ▪ Forms and research plan are very thorough and easy to understand ▪ Student clearly explains all aspects of the project and responds well to all questions

Created by Melissa Potter for the Beaverton-Hillsboro Science Expo, Hillsboro, Oregon, February 2007. Includes criteria from the International Baccalaureate Organization (2001), Massachusetts Department of Education (2006), Oregon Department of Education (2001), and Sneider and Brenninkmeyer (2006).

APPENDIX B

SCIENCE FAIR RUBRIC FOR JUDGES—ENGINEERING

Judge ID:_____Project #: _____

Science Fair Rubric for Judges—Engineering

Please use the following rubric to provide feedback for each student's project. After examining the display board, looking at the research plan, and talking with the student, score the student in each section (Define Problem, Design Solutions, Data Collection, Analysis, and Communication) using the guidelines in the rubric. Also, feel free to circle comments on the rubric that summarize your comments.

Comments (optional):_____

_____

These scores are for feedback only--they do not represent students' final ranking in their category or in "Best of Fair"

Define Problem: _____ **Design Solutions:** _____ **Data Collection:** _____

Analysis: _____ **Communication:** _____

Lab Section **Define the Problem** What to look for: ☐ Description of background material ☐ Description of a practical need or problem	**Score**		
	1	2	3
	▪ Background information missing, incomplete, or only partly relates to the design process ▪ Problem cannot be directly answered, tested, or clearly explained ▪ Criteria for a solution are missing or limited in scope	▪ Background relates to design and accurately uses some supporting literature ▪ Testable problem ▪ Criteria for a solution are clearly identified	▪ Background information is thorough and clear and extensively uses scientific and engineering literature ▪ Problem identifies relevant scientific concepts ▪ Criteria for a solution are thoroughly analyzed (ex., pugh chart) and problem constraints are specified

Judge ID:_____ Project #: _____

Design of Solution* What to look for: ☐ Explore alternatives to answer need or problem ☐ Identification of solution ☐ Creation of a prototype/ model ☐ Testing a prototype *Consider how much help the student received from a mentor.	**Score**		
	1	2	3
	▪ Alternatives are missing or unclear ▪ Solution does not address original need or problem ▪ Prototype/model missing or not adequately designed (i.e., scale incorrect, inappropriate use of materials) ▪ Testing of prototype missing or does not allow for the collection of relevant data	▪ Alternatives are clearly described ▪ Solution aligns with most of the criteria needed to answer need or problem ▪ Prototype/model is adequate to meet the need ▪ Prototype is practical and can be tested to see if it meets the criteria for a successful solution	▪ Alternatives are thoroughly analyzed (ex., pugh chart) ▪ Solution is thoughtfully selected to address the original design constraints ▪ Prototype/model is creatively designed: practical, testable, and meets the criteria and constraints of the problem ▪ Testing of prototype is based on scientific and/or mathematic principles

Lab Section			
Data Collection: What to lookfor: ☐ Data collection procedure is a good test of the design ☐ Tables with labels, titles, and units ☐ Graphs (if appropriate) ☐ Transfor-mation of data (calcu-lations, graphs, etc.) to help explain patterns, trends	**Score**		
	1	2	3
	▪ Data collection not clearly related to a test of the solution to determine if it meets criteria ▪ Tables are not easily understood—missing titles, labels, and/or units ▪ Graphs are unorganized and/or not relevant to the research question (i.e., bar graph when line graph should have been used) ▪ Displays are somewhat appropriate and complete but do not help make interpretations in patterns	▪ Data collection procedure measures adequacy of the design to solve the problem ▪ Tables are clear and well labeled ▪ Graphs are clear and help answer the research question ▪ Patterns and trends in the data are identified and relate to research question	▪ Data collection procedure is thoughtfully designed to test the solution to determine if it meets the criteria and constraints for a successful solution ▪ Tables are thoroughly labeled ▪ Graphs clearly show relationships between variables (and if applicable, supported by the correct and thorough use of statistics) ▪ Patterns and trends are thoroughly discussed and are supported by scientific principles (and if applicable, supported by the correct and thorough use of statistics)

Judge ID:_____Project #: _____

Analyzing What to look for: ☐ Results of the design process ☐ Appropriate-ness of design or technology to solve problem ☐ Evaluation of solutions ☐ Improving the design process	Score		
	1	2	3
	▪ Results presented are not supported by the data ▪ Limited explanation of the appropriateness of design or technology to solve problem ▪ Errors and limitations of the solutions not explained or dealt with in trivial or illogical manner ▪ Ideas for improvement are not based on information gathered or are missing	▪ Clear explanation of results ▪ Clear explanation of the appropriateness of design or technology to solve problem ▪ Clear understanding of the errors and limitations of the solution ▪ Ideas for improvement are based on the information gathered	▪ Clear and thoughtful explanation of the results ▪ Thorough analysis of the appropriateness of the design in comparison with other possible designs ▪ Identification of trade-offs in design decisions ▪ Ideas for improvement are based on the information gathered and address identified weaknesses or limitations
Communi-cation What to look for: ☐ Display board ☐ Forms and Research Plan ☐ Verbal discussion	Score		
	1	2	3
	▪ Display board is unorganized and difficult to understand ▪ Science fair forms and/or the research plan are missing or unorganized ▪ Student struggles to explain some parts of the project (nervousness is okay)	▪ Display board is easy to understand ▪ Forms and research plan are clear and easy to follow ▪ Student clearly explains all aspects of the project	▪ Display board clearly explains all aspects of the project ▪ Forms and research plan are very thorough and easy to understand ▪ Student clearly explains all aspects of the project and responds well to all questions

Created by Melissa Potter for the Beaverton-Hillsboro Science Expo, Hillsboro, Oregon, February 2007. Includes criteria from the International Baccalaureate Organization (2001), Massachusetts Department of Education (2006), Oregon Department of Education (2001), and Sneider and Brenninkmeyer (2006).

APPENDIX C

DEPENDENT AND INDEPENDENT VARIABLE

ZERO ORDER CORRELATION

Dependent and Independent Variable Zero Order Correlation

	Variable					
Variable	1	2	3	4	5	6
1 Quality	1.0					
2 Background	.98	1.0				
3 Method	.99	.99	1.0			
4 Data collection	.98	.98	.99	1.0		
5 Analysis	.99	.99	.99	.98	1.0	
6 Communication	.98	.99	.99	.99	.98	1.0

REFERENCES

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1995). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research, 89*, 197–205.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80,* 139–148.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth-Thomson.

Flowers, C. (2006). Confirmatory factor analysis of scores on the clinical experience rubric: A measure of dispositions for preservice teachers. *Educational and Psychological Measurement, 66*, 478–488.

Gross, P., Goodenough, U., Lerner, L. S., Haack, S., Schwartz, M., Schwartz, R., et al. (2005). *The state of state science standards 2005*. Thomas B. Fordham Institute. Retrieved April 2, 2008, from http://www.edexcellence.net/doc/Science%20Standards.Final.Final.pdf

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan, 80*, 662–666.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*, 1509–1528.

International Baccalaureate Organization. (2001). *Diploma programme: Biology.* Geneva, Switzerland: International Baccalaureate Organization. Retrieved October 10, 2007, from http://occ.ibo.org/ibis/documents/dp/gr4/biology/d4biologui03051e.pdf

International Science and Engineering Fair. (2007a). *Intel ISEF judging guide.* Retrieved April 9, 2008, from http://www.societyforscience.org/isef/isefitems/JudgingGuide.pdf

International Science and Engineering Fair. (2007b). *Intel ISEF 2007 student handbook: Student handbook for precollege science and engineering projects.* Washington, DC: Science Service.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*, 249–267.

Massachusetts Department of Education. (2006). *Massachusetts science and technology framework: Pre-kindergarten—High school standards*. Boston: Author.

McGraw, K., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

McMillan, J., & Schumacher, S. (2006). *Research in education: Evidence-based inquiry* (6th ed.). Boston: Pearson Education.

Messick, S. (1995). Standards of validity and the validity standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5–8.

Miles, J., & Shevlin, M. (2007). *Applying regression and correlation: A guide for students and researchers*. Thousand Oaks, CA: Sage.

National Center for Education Statistics. (2006). NAEP science: Scientific investigations. *The nation's report card*. Retrieved June 10, 2008, from http://nces.ed.gov/nationsreportcard/science/investigation.asp

Novak, J. R., Herman, J. L., & Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of student writing. *Journal of Educational Research, 89*(4), 220–233.

Oregon Department of Education. (2001). *Science standards*. Salem, OR: Author. Retrieved October 8, 2007, from http://www.ode.state.or.us/teachlearn/subjects/science/curriculum/contentstandards.pdf

Oregon Department of Education. (2008). *Science academic content standards revision, K-8 and high school: Request for review*. Salem, OR: Author. Retrieved September 21, 2008, from http://www.ode.state.or.us/teachlearn/subjects/science/curriculum/sciencedraft1.pdf

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*, 347–362.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*, 22–27.

Sneider, C., & Brenninkmeyer, J. (2006). *Achieving technological literacy at the secondary level: A case study from Massachusetts*. Unpublished manuscript, National Center for Technological Literacy, Museum of Science, Boston.

Stokking, K., van der Schaff, M., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Journal, 30*, 93–116.

U.S. Department of Education. (n.d.). *Using data to influence classroom decisions*. Retrieved February 7, 2009, from http://www.ed.gov/teachers/nclbguide/datadriven.pdf

Wiliam, D. (2006, July 11). *Does assessment hinder learning?* Paper presented to the ETS Europe Breakfast Salon, London.