


Spring 7-23-2013

# Impact of Teacher Feedback on the Development of State Issued Scoring Guides for Science Inquiry and Engineering Design Performance Assessments

Timothy Paul Fiser  
*Portland State University*

Let us know how access to this document benefits you.

Follow this and additional works at: [https://pdxscholar.library.pdx.edu/open\\_access\\_etds](https://pdxscholar.library.pdx.edu/open_access_etds)

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Science and Mathematics Education Commons](#)

---

## Recommended Citation

Fiser, Timothy Paul, "Impact of Teacher Feedback on the Development of State Issued Scoring Guides for Science Inquiry and Engineering Design Performance Assessments" (2013). *Dissertations and Theses*. Paper 991.

10.15760/etd.991

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact [pdxscholar@pdx.edu](mailto:pdxscholar@pdx.edu).

Impact of Teacher Feedback on the Development of State Issued Scoring  
Guidelines for Science Inquiry and Engineering Design Performance Assessments

by

Timothy Paul Fiser

A thesis submitted in partial fulfillment of the  
requirements for the degree of

Master of Science Teaching  
In  
General Science

Thesis Committee:  
William Becker, Chair  
Michael Flower  
Emily Saxton

Portland State University  
2013

© 2013 Timothy Paul Fiser

## Abstract

In 2010, Oregon Department of Education (ODE) developed a set of rubrics designed to score a state required performance assessment targeting Science Inquiry (SI) and Engineering Design (ED) skills. During the development of the rubrics, ODE invited six panels of teachers to provide feedback on an early draft of the rubrics. This case study analyzed the teachers' feedback and the revisions of the rubrics to identify the types of feedback teachers offered and how ODE used that feedback to develop the rubrics. The results showed the teachers' feedback focused on defining the skills students were asked to demonstrate and distinguishing levels of student performance. There was clear evidence that the teachers' feedback had a substantial impact on the development of the rubrics. These results suggest that teachers can add substantial value during the development of a state issued assessment tool.

## Dedication

I dedicate this work in memory of my grandmother Virginia Lee Porter Fiser and enduring gratitude to my grandfather Van Eugene Fiser. It was their example, love, and support that helped guide me to a lifelong love of learning.

## Acknowledgements

I would like to express my gratitude and thanks to my advisory committee for their guidance and encouragement as I developed this thesis research. In particular, I would like to thank Dr. William Becker for consistently challenging me as I progressed as a researcher and for making available the funding to begin this research effort. I would also like to thank Dr. Michael Flower for his advice and encouragement which proved invaluable as I approached the final reporting of this research. I especially wish to thank Emily Saxton for her guidance, advice, and persistent encouragement to improve my skills as a writer and as a researcher.

I also wish to thank Dr. Cary Sneider, Jennifer Wells, and Stephanie Wagner at the Portland State University - Center for Science Education for their advice, insight, and support.

I would like to thank the Oregon Department of Education for providing funding and making this research possible.

My thanks also extend to my colleague, Phil Dolan, for his advice in developing data structures and moral support as I worked through my data set. I also wish to thank my colleagues at the Center for Science Education and the Graduate School of Education for their friendship and encouragement during our shared graduate studies.

Finally I wish to express my whole hearted gratitude to my loving wife, my son, my parents, and my siblings whose support, patience, and love kept my spirits high throughout this journey.

## Table of Contents

Abstract .....	i
Dedication .....	ii
Acknowledgements .....	iii
List of Tables .....	vi
List of Figures .....	vii
Introduction .....	1
Methods .....	19
Results of Analysis .....	44
Discussion .....	80
References .....	99
Appendices	
A. Category Matrix for Feedback and Scoring Guide Analysis.....	102
B. Demographic Survey Instrument .....	103
C. Draft V1.5 of Middle School ODE SI/ED Scoring Guide.....	107
D. Draft V1.6 of Middle School ODE SI/ED Scoring Guide.....	111
E. Middle School Scoring Guide for Mark-up.....	115
F. Human Subjects Approval .....	119
G. Semi-Structured Interview Guide.....	120
H. Feedback and Scoring Guide Coding Spreadsheet Example.....	121
I. Feedback and Scoring Guide Analysis Data Structure.....	122



## List of Tables

Table 1: Comparison of rubric design methods between Mertler and Reddy.....	13
Table 2: Scoring Guide Draft Revisions.....	23
Table 3: Organization of six panels by grade level and interest.....	26
Table 4: Number of panelists per panel.....	27
Table 5: Panelist Demographic Data.....	27
Table 6: Titles of Scoring Guide Subsections.....	35
Table 7: Overview of Coding Categories and Subcategories.....	40
Table 8: Number of unique feedback items by category and type.....	61
Table 9: Coded changes in scoring guide drafts V1.6 by type.....	65

## List of Figures

Figure 1: Diagram of Study.....	21
Figure 2: Feedback items by general category.....	62
Figure 3: Distribution of coded feedback items by type.....	63
Figure 4: Distribution of coded feedback items from Panel B3-ED by type.....	64
Figure 5: Distribution of coded changes in V1.6 by type.....	65
Figure 6: Proportionality of change items by category.....	66
Figure 7: Total number of unique feedback items by scoring guide.....	67
Figure 8: Total number of unique change items in V1.6 by Scoring Guide.....	67
Figure 9: Number of changes in V1.6 matching feedback.....	69
Figure 10: Feedback items matching change items as a percentage by panel...	70
Figure 11: Types of feedback with matching changes in V1.6.....	71
Figure 12: Proportion of change items matching feedback by Science Inquiry or Engineering Design.....	72
Figure 13: Categories of feedback without a matching change in draft V1.6.....	72
Figure 14: Proportions of feedback items by score level.....	73
Figure 15: Proportion of change items from all scoring guide versions by score level.....	74
Figure 16: Total number of uniquely coded items by scoring guide draft.....	75
Figure 17: Unique change items in scoring guide drafts V1.7 through V1.9 by type.....	75
Figure 18: Unique change items in scoring guide drafts V1.7 and V1.9 by category and panel.....	76
Figure 19: Unique change items in scoring guide draft V1.8 by category and panel.....	77
Figure 20: Changes in V1.8 by category.....	78
Figure 21: Changes to V1.7 and V1.9 by category.....	78

## Introduction

*Overview.* This is a case study of a period in the development of the 2011-12 Official Scientific Inquiry/Engineering Design Scoring Guides for the state of Oregon. During the period of time under study there were several events relating to the development of the scoring guides that were of particular interest. The primary event understudy in this research was the work done by several panels of experienced in-service teachers which were gathered to evaluate an early draft version of the 2011-12 Official Scientific Inquiry/Engineering Design Scoring Guides, henceforth referred to as the SI/ED Scoring Guides, and to report back to the Oregon Department of Education (ODE) the panels' feedback and recommendations for changes for the SI/ED Scoring Guides.

This study conducted a detailed line by line analysis of the feedback the Teacher Evaluation Panels offered to ODE and the changes observed in several draft revisions of the SI/ED Scoring Guides. These analyses were conducted to answer the two research questions that guided this study.

1. What types of feedback did the panels of experienced teachers offer ODE for the revision of the SI/ED scoring guides?
2. How did ODE utilize the feedback the teachers offered to revise the SI/ED Scoring Guides?

The SI/ED scoring guides were designed to score evidence of a student's proficiency as demonstrated through a work sample performance assessment of Scientific Inquiry or Engineering Design. The results of this study showed that the feedback offered by the Teacher Evaluation Panels was focused on recommending changes that clarified or refined the fundamental functions of the

scoring guides and that the feedback was used to make changes to the scoring guides likewise clarifying and refining the fundamental functions of the scoring guides and arguably improving the scoring guides.

*Background.* In 2009, Oregon adopted a new set of Science Content Standards. Included in these updated standards was an overhaul of the Scientific Inquiry (SI) standards, originally implemented in 2002, and the addition of new Engineering Design (ED) standards. The content standards for SI and ED are intended to incorporate a conceptual understanding of the nature of science and processes of engineering with the commonly canonized science curriculum subject areas: Physical Science, Life Science, and Earth and Space Sciences (Kleckner, 2009).

There were two main components to the Oregon science assessment strategy to assess student learning with respect to the science content standards revised in 2009. The first was the long standing and updated standardized multiple choice exam called the Oregon Assessment of Knowledge and Skills, commonly referred to as OAKS (Vanderwall, 2011). The second method was relatively new and was still in the process of being implemented, that was a local performance assessment of Scientific Inquiry and soon to be added Engineering Design.

In June of 2008, it became Oregon law to include the requirement that Oregon public schools use local performance assessments to evaluate their students' progress. For grades 3<sup>rd</sup> through 8<sup>th</sup>, at least one performance assessment in Science Inquiry, or Engineering Design, was to be required per

year using (Test Administration Manual - ODE, 2009). Starting in 2012, it was expected to be a graduation requirement for high school students to demonstrate proficiency in either one SI or ED performance assessment during their high school career (Phillips, 2009).

It was not required by state policy that schools use the Student Work Sample performance assessment method supported by ODE. The language of the law allowed for another locally developed performance assessment be used for the local performance assessment (Test Administration Manual - ODE, 2009). However, if the Student Work Sample method was chosen, it must be scored using the official state scoring guides (Test Administration Manual - ODE, 2009) which were the subject of this case study. The intended requirements that these scoring guides would be used as a gate for grade level promotion and high school graduation considerably raised the stakes for the SI/ED Scoring Guides.

#### *Standards Reform and Skills Standards*

This study was conducted during a process of standards-based reform and assessment reform in the state of Oregon. ODE has been actively pursuing standards-based reform for approximately 20 years (Svicarovich & Kirk, 2009). The implementation of these reforms involves many elements and stages planned out over time (Vanderwall, 2011). Among these stages were the publication of a new set of state wide science content standards in 2009 and the development of new assessment tools to assess student learning with respect to these new science content standards. Among the assessment tools being developed were the SI/ED Scoring Guides, which were the subject of this study.

*Standards-Based Reform.* In a peer reviewed editorial article on the subject of standards-based education reform, Thompson (2001) distinguished what he called 'authentic standards-based reform' as an attempt to address the quality of education in a forthright manner through making clear and public the high expectations and educational goals for which teachers and students were to be held accountable. This is in stark contrast to test-based reform, which has been confused by some with authentic standards-based reform (Thompson, 2001), and has been interpreted as having the opposite effect of authentic standards-based reform in that test-based reforms reduced education to a focus on ensuring students achieve minimum test scores in order to avoid negative political consequences (Faladi & Robinson, 2000). In test-based reform, the test themselves become the standards to which teachers teach (Faladi & Robinson, 2000, Thompson, 2001).

In Oregon, the OAKS exams are standardized tests and remain a part of the ODE science assessment program to assess student knowledge (Vanderwall, 2011). However, the OAKS exams are explicitly not the education standards for the state of Oregon. ODE published education standards for the state, which are explicitly the standards. Further including local performance assessments in the assessment battery for the state is a move away from relying solely on the OAKS assessments and for the statewide assessment of science in the Oregon. The addition of the local performance assessments for science would then prevent a 'single indicator assessment', such as the OAKS tests, which could then narrow the scope of curriculum and thereby reduce the role of

teaching to that of test preparation (Thompson, 2001). And the use of the broader range of assessment techniques helps to ensure a rich, contextual, authentic learning and teaching experience, which is the aim of standards-based education reform (Wiggins, 1998).

*Skills Standards.* Science Inquiry as a skills and process standard was included for the first time in the Oregon Science Standards in 2002. The Oregon Core Science Standards were expanded to include Engineering Design along with Science Inquiry in the 2009 Oregon Science Core and Content Standards.

The purpose for including the SI and ED skills standards in the Oregon Science Standards was to ensure that students understand the process skills and concepts that are characteristic of the practice of science and engineering (Vanderwall, 2011). ODE defined Scientific Inquiry (SI) and Engineering Design (ED) as follows:

Scientific Inquiry is a systematic process that includes proposing testable hypotheses, collecting, analyzing, and interpreting data to produce evidence-based explanations and new explorations.  
Engineering Design is a process of formulating problem statements, identifying criteria and constraints, testing solutions, and incorporating modifications based on test data and communicating the recommendations. (Science Assessment – ODE. 2011, p.5)

The knowledge and skills represented in these standards extend beyond the content of science knowledge already attained in the history of science and engineering. Rather the knowledge and skills in these core standards pertain to how new scientific knowledge is obtained, communicated, and understood, in the case of SI, and in the case of ED, how science and technology are applied to

solve problems in society. The SI and ED standards push the curriculum to extend beyond rote knowledge and traditional problem solving exercises and to develop students' understanding of how to apply scientific knowledge in ways that are authentic to the ways scientists and engineers work.

The effects of using a curriculum that is specifically focused on developing Scientific Inquiry knowledge and skills was shown in experimental studies to have a significant effect on the ability of students to understand and execute scientific reasoning (Turner, 2008). The SI Scoring Guides within this study are intended to similarly measure students' developing understanding and reasoning with respect to the process of conducting a scientific investigation.

The focus on Engineering Design in curriculum has been a growing national concern (Faladi & Robinson, 2000; National Research Council, 2011). There is a recognized national economic importance of engineering and the interdependence between Engineering and Science add to the urgency to include engineering in science education (National Research Council, 2011).

There are two main goals driving the national movement to include ED along with SI in standards reform. The first is to simply increase the number of students that pursue upper level science classes. Second is to increase the number of students who have taken upper division science courses to then pursue careers in engineering (National Research Council, 2011). Faladi and Robinson (2000) reported "many high school teachers and students (were) largely ignorant of what engineers do" (p. 3). Through explicit instruction in ED, students would be better equipped to evaluate engineering as a career



option. At the time this research was being conducted, the National Core Standards were still articulating the inclusion of engineering into the science standards for K-12 education. By including ED in the science standards and articulating the assessments to measure learning gains in this domain, the science education reform in Oregon is ahead of the national curve with respect to K-12 engineering education (National Research Council, 2011).

*Testing, Performance Assessments, and Rubrics.* Standardized tests are generally designed to measure a minimum competency in student learning or growth (Faladi & Robinson, 2000) which starkly contrasts with the degree or extent of understanding and ability for which authentic standards-based education reform is designed to enhance (Thompson, 2001). This makes standardized tests poorly suited to assess the higher order knowledge and skills emphasized in the Science Inquiry and Engineering Design standards within the 2009 Oregon Science Standards. For decades, there has been a strong interest in developing assessment alternatives to standardized tests and a desired progress toward performance based assessments. (Shavelson, Baxter, & Pine, 1993).

Baker and O'Neil (1994) define performance assessment as follows:

Performance assessment is student testing characterized by constructed responses, long-term engagement in project-like tasks. Student performances, either on-the-spot, hands-on behavior, or their products, such as extended reports, or works of art, are judged by experts to determine their quality (pg 2).

In the context of the Oregon local performance science assessment, the performance assessments give students the opportunity to demonstrate their

proficiency and understanding of science in a more authentic and direct manner (Wiggins, 1995; Linn, 1994). Some of the hallmarks of performance assessments include complex learning, higher order thinking, active response, complex tasks, and a significant investment of time spent on the tasks (Baker & O'Neil, 1994). ODE stated that local science performance assessments are intended to “evaluate the application of students’ knowledge and skills” (Vanderwall, 2009, p.2). The assessment of students' ability to apply SI and ED knowledge and skills outlined in the 2009 Science Content Standards is at the center of the local performance assessment and the target of the SI/ED Scoring Guides.

Research efforts to improve upon assessment techniques or tools, or to report on improved assessment techniques or tools, is a rich domain in the literature. (Liu, Lee, Hoffsetter, & Linn, 2008; Day & Matthews, 2008; Hammond, 2003; Jonnson & Svingby, 2007; Shavelson et al., 1992; Baker & O'Neil, 1994, Mertler, 2001; Mertler 2011; Reddy, 2011). These studies range from addressing validity in general (Messick, S., 1994; Mertler, 2001, 2011) to design and/or evaluation of new assessment tools (Reddy, 2011; Liu et al., 2008; Day & Matthews, 2008.)

There is a growing preponderance of evidence supporting the use of performance assessments. It has been shown that performance assessments can improve instruction (Hammond & Adamson, 2010). The use of rubrics to score work was found to be beneficial to improve student performance as well as increase the validity and reliability of assessments (Jonnson & Svingby, 2007). With certain caveats understood, performance assessments have been

shown to reduce the effects of racial biases in assessment (Shavelson et al., 1993; Baker & O'Neil, 1994).

Not all of the research points to using performance assessments to assess students' understandings and skills working with SI or ED. Some skepticism was reported regarding how well suited a performance assessment approach can be for a wide spread assessment tool (Day & Matthews, 2008), such as the state wide assessments intended by Oregon's local science performance assessments and the accompanying SI/ED Scoring Guides.

Day and Matthews (2008) offered an analysis of an assessment tool that was a part of the New York State Education Department's battery of assessments which, like the Oregon local performance assessment and SI/ED Scoring Guides, was designed to specifically assess Science Inquiry knowledge and skills (Day & Matthews, 2008). Unlike the Oregon work sample performance assessment the NYSED assessment was a more traditional paper and pencil exam with prompts and student responses. The results reported that the NYSED assessment tool inadequately address the breadth of the inquiry standards it was intended to measure.

However, contrary to the direction Oregon is pursuing, Day and Matthews (2008) recommended improving the standardized test instrument in the hope of improving its performance. While it was their opinion that performance assessments could be better suited to meaningfully measure SI skills, they were skeptical that large scale performance assessments could be a practical solution (Day & Matthews, 2008).

Some research was found of the development of assessment tools that targeted higher order thinking skills, such as those that would be expected in the Oregon SI or ED assessments. Though these assessment tools were not explicitly performance assessments, the research reported on assessment tools that may be more easily adapted to a wide spread application, such as a state wide assessment, and may prove to be better suited for assessing these higher ordered thinking skills than the standardized multiple choice tests. Liu et al., (2008) reported on an NSF funded hybrid assessment tool that employed open response items for the students to answer and was scored with a logic decision tree guiding the rater to higher or lower orders on the rubric based on the students' responses. The results of this study were generally positive and there were multiple similarities between the hybrid assessment and the Oregon assessment: both targeted science inquiry, both utilized rubrics and both were still under development and evaluation. While the initial results for this hybrid assessment tool were largely positive, some concern was expressed about the training needed due to the complexity of the assessment scoring process.

The approach taken by Oregon to use a performance assessment for the wide spread assessment of high order SI or ED knowledge and skills is somewhat of a novel approach. It remains to be seen whether Day and Matthews (2008) opinion about the feasibility of large scale assessment made through performance assessment is correct. The success of these performance assessments in Oregon could be significantly affected by the quality of the SI/ED Scoring Guides as well as the willingness for teachers to adopt and use the

assessment tool. The quality of the scoring guides and possibly how successful the implementation of the scoring guides are discussed in this study.

*Rubric Development.* Since this case study collected data on several stages of development of the SI/ED Scoring Guides, some description on rubric design and development is appropriate here.

The SI/ED Scoring Guides are analytic rubrics. Analytic rubrics are divided into several distinct sections, score a student's work against some prescribed criteria for each section in the rubric, and then sum the scores of the sections to compute a final score (Mertler, 2001). See Appendix D for an example of an SI/ED Scoring Guide. Each scoring guide within the set of SI/ED Scoring Guides has four distinct sections. In the case of SI Scoring Guides, the four sections are 'Framing the Investigation', 'Designing the Investigation', 'Collecting and Presenting Data', and 'Analyzing and Interpreting Results'. Each one of these sections could be considered an analytic rubric in itself as there are multiple criteria for each score level in the scoring guide.

The opposite of analytic rubrics are holistic rubrics which have only one section and the criteria for the score levels within the holistic rubrics are typically broadly defined allowing for a wider variety of student responses. This contrasts with the analytic rubric in two important ways. First, the analytic rubric has a much higher order of complexity when it comes to evaluating and analyzing the tool. However, in trade, the specificity of the criteria and the organizational structure allows for clearer interpretation of the expectations expressed in the criteria. The trade-off for this clarity is that analytic rubrics are more likely to be

prescriptive with respect to the type of evidence it can score (Mertler, 2001). The criticism of this is that it can limit creativity on the part of the student. However, as communicated by the scoring guides, there are multiple elements that each student will be expected to include in their performance assessment. See the criteria in Appendix D for examples of these required elements.

Reddy (2011) reported on a pilot study of the development of rubrics with the intent of improving assessment outcomes for students. Within his study, he proposed a method involving eight steps to develop rubrics to be used for program assessment. One of the purposes of the state wide use of the OAKS assessments is to evaluate science instruction at the state and district level (Vanderwall, 2009). Though it is speculative on the part of the researcher, the state wide use of the local science performance assessment could also be useful to programmatically evaluate science instruction within the state provided sufficient care was given to validate the assessments.

Mertler (2001) suggested seven steps to follow when developing a scoring guide. The eight steps to develop a rubric recommended by Reddy (2011) appear to borrow heavily from the Mertler (2001) though when compared side by side there are some notable differences. These steps are outlined in Table 1.

**Table 1: Comparison of rubric design methods between Mertler and Reddy.**

Mertler (2001)	Reddy (2011)
1. Examine the learning objectives of the task.	1. Identify the learning objectives to be served by the use of the assessment method and which lead to the identification of qualities (criteria) that need to be displayed in a student's work to demonstrate proficient performance
2. Specify observable attributes that will demonstrate their proficiency.	2. Identify levels of performance for each of the criteria.
3. Brainstorm characteristics that describe the above attributes.	3. Develop separate descriptive scoring schemes for each evaluation level and criteria.
4. Write a thorough description for excellent and poor work for each attribute.	4. Obtain feedback on the rubrics developed.
5. Describe other levels of proficiency on the scale.	5. Revise the rubrics based on feedback from primary stakeholders.
6. Collect student work samples that are exemplary of the scale levels.	6. Test the reliability and validity of the rubrics.
7. Revise as necessary.	7. Pilot test of the rubrics.
	8. Use the results of the pilot test to improve the rubrics.

Steps 1 and 2 in Mertler's method were merged into the first step in Reddy's method but these methods both agree that the first steps are to identify the objectives the rubric is to measure and then to collect 'attributes' or 'criteria' that provide evidence regarding how well the student met the desired objective. These steps will be notable again in the methods and results section below.

Where Mertler and Reddy differ, starting at Reddy's step 4, is Reddy's recommendation to obtain feedback on the rubric from primary stake holders and then to revise the rubric on the basis of that feedback (Reddy, 2011). These are

the very steps captured in this case study. The results of the analysis herein will report on the types of feedback that were collected from the stake holders, in this case in-service teachers that are expected to be using the scoring guides within the following year. And then through the analysis, conclusions will be drawn describing the impact the teachers' feedback had on the scoring guides.

### *Teachers and Policy Initiatives*

*Teacher Involvement.* In this study, teachers were asked to participate in the development process of a state assessment tool. Very little research was found that reported on teachers providing feedback to policy makers, especially regarding teacher feedback to a state department of education. The lack of research literature in this area was also reported by Reddy (2011). However, several anecdotal accounts were found where teacher input was received and accepted by policy makers. One example was in the state of Montana during the development of a state-wide policy initiative. The Office of Public Instruction acquired feedback through multiple means from education professionals, including in-service teachers. Based on the non-academic article, the feedback appeared to have been utilized by the state as it continued to develop its policies (Barlow, 2009).

Another example was found in an article recently published anonymously in the magazine *American Teacher*. In this case, in-service teachers took the initiative to review a draft of proposed Mathematics and Language Arts



Standards and then offered feedback to the Council of Chief State School Officers. In the teachers' feedback, they noted several omissions of content details that were immediately apparent to them as active practitioners, but seem to have been missed by the policy makers who had not likely been in a classroom for many years (anonymous, 2009).

Research literature was found from several additional sources where it was recommended that in-service teachers should be included in the process of developing wide scale assessments. In a scathing peer reviewed editorial concerning the state of the industry of high stakes standardized assessments, Gallagher (2002) questioned the validity of any assessment that was developed in secret by "remote experts". He went on to ridicule the spectacle of some industrial assessment developers for recruiting teachers to participate in the development of assessments, but in the end these teachers were brought in effectively as a public relations ruse. On the contrary, Gallagher (2002) recommended that teachers be recognized as professional assessors of student work and that assessments should enable the teachers to do the work of assessing their students within the context of the classrooms. Researchers evaluating another statewide assessment tool recommended that panels of teachers be recruited to evaluate the SI items in the exam with the lab experiences the students received in the classroom to better align the exam to the instructional experiences the students were getting in the classroom (Day & Matthews, 2008).

*Teacher affect and morale.* Further, regardless of the reported benefit lost by not including teachers in the development of these assessment tools, there are other costs to excluding teachers from the process. The new policy of the local science performance assessment, and the accompanying SI/ED Scoring Guides represent, are education reforms that were generated by the ODE and the state legislature.

Despite good intentions, when high stakes top-down education reform occurs, it has been shown there are can be profound and sometimes very counterproductive consequences in the classroom and beyond. Valli and Buese (2007) conducted a longitudinal study of the impact of the implementation of a series of assessment initiatives on elementary teachers. The study concluded that in the midst of the assessment reforms the teachers had difficulty reconciling their practice to the increased roles the teachers were expected to fill as a result of the reform, and the teachers experienced a deterioration of professional wellbeing. In addition to the teachers' morale, a deterioration of pedagogical practices and relationships with the students were also identified in the study. These latter effects were noted more frequently in high needs, title-1 schools (Valli & Buese, 2007).

Similar deteriorations in teacher morale in the face of education reform were identified in a study of teacher beliefs (Lumpe, Hanley, & Czerniak, 2000). The result of reduced morale was a lower probability that the reform would be adopted by the affected teachers. Another study looking at the negative impact

top down reforms had on teachers and the adoption of reforms found that in addition to a low rate of teachers adopting the new reforms into their practice, there can also be a heavy toll taken on the culture of the school and district with a high turn-over rate for the leadership and teachers that were open to the reform as a result of being pushed out of the school by disaffected teachers (Olsen & Sexton, 2009).

The research offered some suggestions to remedy or to avoid these observed negative consequences that can occur with top down education reform. The implementation of the reform should be carried out while maintaining respect for the teachers that would be affected by the reform (Olsen & Sexton, 2009). Before and during the implementation of a reform it was recommended that the attitudes of the affected teachers be assessed and that professional development opportunities be offered to address issues that might threaten the success of the reform (Lumpe et al., 2000). Including the teachers in the policy decision making or development process, as well as being aware of the amount of time the different reforms may take before the teachers are comfortably ready for the next phase of reform, was strongly recommended by Valli & Buese (2007).

Several of the concerns and recommended practices from the above studies appear to have been taken into consideration and acted upon by ODE while it continues to develop and improve the standards-based educational system for the state. As ODE continues to refine the standards-based policies, changes to the policies have been scheduled in a methodical and forecasted

manner (ODE Science Standards Adoption, 2009) so as to inform practitioners as well as other stake holders of what policy changes are expected and when to expect them. Further, and more specifically to the context of this study, through professional development workshops scheduled prior to the official release of the SI/ED Scoring Guides, ODE had an opportunity collect data about teachers' attitudes and to address teachers' knowledge and beliefs about the 2010 SI/ED Scoring Guides. Finally, by including teachers in the process of developing this state wide assessment tool, that is by supporting the Teacher Evaluation Panels in order to collect in-service teachers' feedback concerning the new scoring guides, ODE is clearly answering the call to include teachers in the education reform process.

This case study looks at the effect several panels of experienced teachers had on the development of the SI/ED Scoring Guides. The teachers' feedback was analyzed in detail as were the draft versions of the scoring guides as they progressed from early versions to late versions. The results show strong evidence of the contribution teachers made working toward the end product.

## Methods

### *Overview*

This thesis research was a case study of the development of the 2011 Oregon Science Inquiry and Engineering Design Scoring Guides which were developed to score the local performance assessment of elementary through secondary student work samples targeted to demonstrate proficiency in Science Inquiry (SI) or Engineering Design (ED) process knowledge and skills. Specifically, this study focused on the how the scoring guides changed during the development process following the input of 6 panels of experienced in-service teachers. This case study set out to answer the following questions:

1. What types of feedback and recommended changes did the panels of experienced teachers offer the ODE?
2. How did ODE use this feedback during the continued development of the scoring guides?

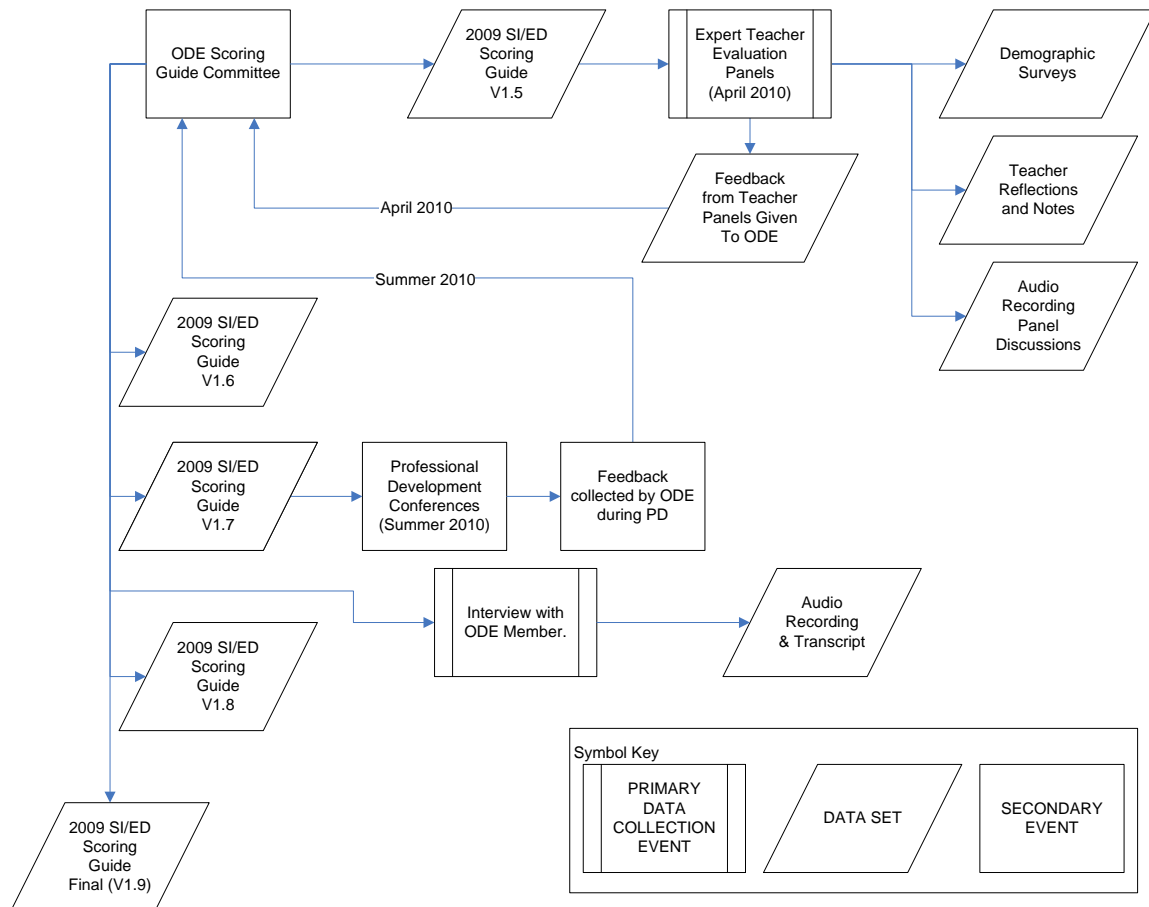
To answer these questions a mixed method case study approach was taken. Borrowing from the typology of Thomas (2011) this case study was a single key case study describing the diachronic development of the scoring guides through several iterations to explore the impact, and possible value of, the input offered by experienced teachers toward the development of those scoring guides.

### *Case Under Study.*

In April of 2010, six panels of experienced in-service teachers were convened to review and remark upon draft versions of the 2010 Oregon Science Inquiry and Engineering Design Scoring Guides. The six panels were organized by grade level and the panelists' interest in science inquiry or engineering design. Each

panel worked with the corresponding scoring guide. See Table 2. The panels were convened all day Friday April 20th, 2010 and for a half day on Saturday April 21st, 2010. This study collected data from these panels and the subsequent changes made to the scoring guides.

The diagram in Figure 1 below shows the sequence of several events and processes through which data were collected for this case study. The data collected came from two types of events: The first type of event is the primary event that was structured by the researchers to provide the opportunity for the teacher panels, the participants, to evaluate and provide feedback for the development of the SI and ED Scoring Guides. This event was designed by the researchers in order to facilitate 1) the work of the panels and 2) data collection for this study. As denoted in the diagram below, there were several secondary events from which data was collected for this study. These events and processes were conducted by ODE independently of the researchers and were external to the control of the researchers. However, these events were important sources of data, which pertained directly to the questions this study set out to answer and provided further insight into data collected from the Teacher Evaluation Panels in the primary data collection event.



**Figure 1: Diagram of Study**

*Primary Events.* The first primary data collection event was organized by the researchers to provide a focused atmosphere for the Teacher Evaluation Panels to consider and discuss the scoring guides in detail. The explicit goal for the panels was for each panel to generate a report reflecting the opinions and/or recommendations for changes to the scoring guide that panel evaluated. These feedback reports were requested by ODE for the ODE Science Content and Assessment Panels to consider in their continued revision of the scoring guides.

The second primary event was a telephone interview of a primary member of the ODE Science Content and Assessment staff member who was party to the meetings of the ODE Science Content and Assessment Panels and the internal processes within ODE. The interview was conducted on August 30th, 2010, after ODE had received the feedback of the Teacher Evaluation Panels and released a new draft version of the scoring guides. It is notable here that the interview occurred after the professional development events, discussed below, were conducted and a second collection of teacher feedback was gathered and SI/ED Scoring Guide draft V1.8 had been released.

*Secondary Events.* These events were outside the control of the researchers. For the most part these events were conducted by ODE. The first event shown in Figure 1 was a secondary event, however it was the catalyst for this study. This was the development and release of an early draft version (V1.5) of the SI and ED Scoring Guides by the authoring ODE Science Content and Assessment Panels. These draft versions of the scoring guides along with some supporting documents were then provided to the Teacher Evaluation Panels.

Throughout the remainder of the year, the ODE Science Content and Assessment Panels released several subsequent draft versions of the scoring guides. During analysis a short hand tag, for example V1.6, was ascribed to each draft version. This tag was based on a draft version-tracking scheme that ODE partially employed during the development of the scoring guides. These draft versions, V1.6 through V1.9, corresponded in timing with several events which occurred over the summer. These events provided a context with which



the revisions of the developing scoring guides could be scrutinized. Table 2 shows an overview of these revisions and the corresponding events.

**Table 2: Scoring Guide Draft Revisions**

Draft Name	Draft Release Date	Event
V1.5	April 2010	Draft offered to Teacher Evaluation Panels for review and feedback.
V1.6	5/12/2010	Draft following feedback from Teacher Evaluation Panels
V1.7	6/1/2010	Draft used for Summertime Professional Development events
V1.8	8/25/2010	Draft following Summertime Professional Development events during which additional feedback was collected from participating teachers.
V1.9	12/16/2010	Draft recommended for school districts to Beta test scoring guides

It is worth noting here that the scoring guides are organized into benchmark levels for clusters of grade levels. The scoring guides included in this study were for three benchmarks: benchmark 2 (B2) includes upper elementary school, benchmark 3 (B3) includes middle school, and the high school benchmark (HS). See Table 3 for exact grade levels. Any given release of a draft version of the scoring guide included scoring guides for each benchmark. For example, draft version V1.6 was frozen on May 12th. This draft version included each of the six scoring guides included in this study: SI and ED scoring guides for benchmarks B2, B3, and HS.

During the summer there were several professional development (PD) meetings conducted around the state. From information gathered through the interview with the ODE staff member it was learned that Draft V1.7 was the

released by the ODE Science Content and Assessment Panels prior to the PD events and was the draft version used during these PD events. During the summertime PD events, ODE introduced the scoring guides to teachers around the state and collected additional feedback on the scoring guides from attending teachers as well as through a survey that was posted online. How the feedback was prompted and collected was outside the purview of this study. However, based on the interview with the ODE staff member, that feedback resulted in the next draft of the scoring guides, V1.8. This draft version provided additional data on how the scoring guides were potentially affected by teacher feedback.

The final versions of the scoring guides were released in December, 2010. These versions of the scoring guides were intended to be used for beta testing around the state according to the interview with the ODE staff member. Though not consistently labeled as such by ODE, this final draft was denoted in the analysis section as V1.9.

*Data Sources:* The central data collected in this study were the feedback documents submitted by the Teacher Evaluation Panels and the 5 draft versions of the SI and ED Scoring Guides created by the ODE Science Content and Assessment Panels.

To gain additional insight into the participants' perspective and background, a survey was collected. Audio recordings of discussions the panelists had while they wrote the panels' feedback reports provide additional insight into the processes and opinions of the teacher panels. Finally, an

interview with the principle member of ODE discussing the development of the scoring guides provided one more perspective to the data collected for this study.

*Participants.* In April of 2010, six panels of experienced in-service teachers were convened to review and remark on draft versions of the 2011 Oregon Science Inquiry and Engineering Design Scoring Guides. The costs incurred for substitute teacher time for the attending teachers were shared as part of the partnership between Portland State University (PSU) and ODE. The participants of this study were the members of the Teacher Evaluation Panels, which were convened in April 2010 and a principle representative of ODE who was interviewed separately.

The in-service teachers were identified and then recruited to participate in this study by Teachers On Special Assignment (TOSAs) who were also working with the Center for Science Education (CSE) at PSU as part of the TOSAs' special assignment. The TOSAs had roles, which were funded as fractions of Full Time Equivalent (FTE) as follows: 0.25FTE funded through CSE of PSU, 0.25FTE funded through the teachers' home districts. These 0.5FTE equivalent roles were assigned to work on partner projects between PSU and the TOSA's school districts. The remaining 0.5FTE roles for most TOSAs were most commonly working as classroom teachers or as teacher mentors in their home districts.

The primary criteria for identifying a potential panelist was the prospective panelists' professional teaching experience using performance assessments of Science Inquiry work samples and the 2002 Science Inquiry scoring guides

and/or their knowledge and interest in teaching Engineering Design. The panelists who participated in this study were selected for recruitment based primarily on a TOSA's professional experience working with, and knowledge of, the recruited teachers' work experience.

There were six panels, one for each scoring guide included within this study. Each panel was designated as either Science Inquiry (SI) or Engineering Design (ED), three panels each. Each group of SI or ED panels was delineated by the benchmark grade levels the scoring guides were to be used. The teacher participants worked in panels within the benchmark at which teachers taught: Elementary, Middle School, or High School. Table 3 below provides a key to the short-hand nomenclature used to identify each panel. After the in-service teachers agreed to participate in the study, the teacher panelists were given the choice of which panel, for their grade level, they would prefer to participate: SI or ED.

**Table 3: Organization of six panels by grade level and interest.**

Grade Level	Science Inquiry (SI)	Engineering Design (ED)
Elementary Benchmark 2 Grades 3, 4, 5	B2-SI	B2-ED
Middle School Benchmark 3 Grades 6, 7, 8	B3-SI	B3-ED
High School Grades 9, 10, 11, 12	HS-SI	HS-ED

The original hope was to recruit a sufficient number of teachers for each panel such that there would be four teachers per panel. However, due to time constraints before the panels were to be convened and immediate access to teachers who were available or interested in participating in the panels, the number of teachers per panel was not evenly distributed. The actual distribution of panelists among the panels was recorded in Table 4.

**Table 4: Number of panelists per panel.**

	B2-SI	B2-ED	B3-SI	B3-ED	HS-SI	HS-ED
Number of Panelists	3	3	5	5	5	3

*The Panelists:* There were 24 panelists involved in the study. Of those 24, 23 completed and returned the demographic survey. See Table 5.

**Table 5: Panelist Demographic Data**

Category	Participants (%)											
Gender	78	Female	22	Male								
Ethnicity	87	Caucasian	9	Hispanic	4	Amer-Indian						
Years Teaching	26	9-11 years	22	> 15 years	22	6-8 years	17	3-5 years				
Highest Degree Attained	57	MA or MS	26	Multiple MA or MS	13	BA or BS	4	Ph.D. or Ed.D				
Under-graduate Majors	56	Science	22	Other Disciplines	13	Science Education & Science	9	Elementary Education				
Graduate Majors	25	Middle School Education	21	Elementary Education	17	Science Education & Science	13	Science Education	8	Science	4	Other Disciplines

While 83% of the teachers had more than 5 years of teaching experience at the high school level, none of the panelists had less than 3 years of teaching experience. One of the teachers counted in the '3-5 years of teaching

experience' category noted in the margins of the survey that the participant had taught for over 20 years at the college level. Likewise the panelists were well educated. All respondents had earned a bachelor's degree and 87% of respondents had earned at least one graduate degree.

*ODE Science Content and Assessment Panels.* Though members of these panels are not participants in this study, a brief description of the types of people in these panels is relevant here. Based on the interview with the principle member of ODE, the ODE Science Content and Assessment Panels are comprised a wide variety of professions. Some of the members are university professors, some are professionals in the private sector, such as scientists or engineers, some members are in-service teachers from different grade levels. That some teachers are already on the ODE Science Content and Assessment Panels means that teachers were already involved in the development of the SI/ED Scoring Guides with or without the Teacher Evaluation Panels. However, it is unknown how much of a voice or impact the teachers in the ODE Science Content and Assessment Panels had during the development of the SI/ED Scoring Guides.

### *Instruments*

There were no existing instruments identified that fit the exact purpose of this research so the instruments used in this research were developed within this thesis project. There were three instruments used for this study: a demographic survey discussed above, a semi-structured interview, and a Categorization Matrix

used to code the feedback from the teacher panels and the changes identified in the evolving draft versions of the scoring guides.

The demographic portion of the survey used in this study, see Appendix B, was based on the Surveys of Enacted Curriculum (SEC) Science Teacher Survey for Grades K-12. This survey is a well-established survey developed by the Council of Chief State School Officers (Blank, Porter, & Smithson, 2001). The SEC is a self-reporting tool designed to collect reliable and comparable data to aid the planning of instructional improvements for teachers, administrators, and policy makers. The SEC was validated using a comparative analysis sampling technique in drawing conclusions from the following data: teachers' responses to the survey, instructional logs, observations by outside researchers, and student survey responses. The end result of the validation analysis was a consensus of high reliability (Blank et al., 2001).

The complete SEC survey is extensive and would have required a significant amount of time for panelists to complete. In addition to general demographic information, the SEC survey would have also inquired about teachers' instructional practice: active learning in science, content in the classroom, assessment strategies, utilization of technology, teacher preparation, professional development, influences of policy on teaching practice, and alignment of instruction to state assessments (Blank, 2005). Not all of the items covered in the survey were specifically informative for this study and there were areas of interest not covered within the SEC survey that were more germane to this study. In the interest of efficiently acquiring data considered to be most likely

germane to this study, the SEC survey was used as a guide and in some cases provided the exact wording for some of the items used in the survey employed in herein.

*Semi-structured interview.* A semi-structured interview was conducted with a principle member of ODE. The ODE member was the Science Assessment Specialist for the state of Oregon. Her duties included "managing, providing development and maintenance of (Oregon's) current content standards, (Oregon's) newly adopted standards and the assessment of the content standards". This interview targeted evidence that would 1) confirm or disconfirm whether the feedback offered by the teacher panels was considered or used to make revisions when ODE drafted version 1.6 of the scoring guides, 2) to further gain insight into the decision making process that went into considering the teachers' feedback and 3) obtain evidence of the value the ODE Science Content and Assessment Panels perceived in the teachers' feedback. The questions written for the interview guide concentrated on the development of the scoring guide drafts V1.5 and V1.6 including the utilization, or non-utilization, of in-service teacher written recommendations based on draft V1.5 from the evaluation panels held April 2010. See Appendix G. The interview was conducted after the summertime PD meetings and would yield data beyond the originally proposed purpose.

*Categorization Matrix.* In order to answer the first research question regarding what type of feedback would the Teacher Evaluation Panels offer, it



was necessary to develop a method to identify the types of feedback the teacher panels offered. To answer the second research question concerning how the Teacher Evaluation Panels' feedback might be used by ODE, it was necessary to develop a method to collect evidence that might link the feedback from the Teacher Evaluation Panels to the changes that made in the scoring guides. A Categorization Matrix was developed to facilitate these necessary analyses. See Appendix A. The details describing the development of the Categorization Matrix are described more fully in subsection *Developing the Categorization Matrix* in the 'Data Analysis' section below. This analysis tool was developed ad hoc during the analysis of the feedback and scoring guide data sets.

#### *Procedure*

The data for this case study were collected over the course of several months, between April and December of 2010. The first primary data collection event was the day and a half day meeting of the 6 Teacher Evaluation Panels. The details of that event are described in the 'Event Structure' subsection below.

The subsequent draft versions of the scoring guides (V1.6 through V1.9) comprised a considerable block of data that was received from ODE as the drafts were completed and made available for external review. The second primary data collection event was an interview with a principle member of ODE was conducted during the summer of 2010 after the feedback from the Teacher Evaluation Panels had been reviewed by the ODE Science Content and Assessment Panels and two draft versions (V1.6 and V1.7) of the scoring guides had been released by ODE. The final two drafts of the scoring guides included in

this study, V1.8 and V1.9, were collected and added to the data to be analyzed using the Categorization Matrix mentioned above.

*Event Structure.* During the primary data collection event 6 panels of experienced in-service teachers were convened on Portland State University (PSU) campus for one and a half days: Friday April 23<sup>rd</sup> from 8am to 4pm and Saturday April 24<sup>th</sup> from 8am until noon. The event was organized and facilitated by Emily Saxton and Timothy Fiser.

The first day the panels met included a brief presentation by a principle member of ODE to orient the panelists to the scoring guide development process, the latest draft of the SI/ED Scoring Guides, and the ODE work sample requirements. Following the presentation the teachers were instructed to separate into the panels to begin their work.

Each panelist received a folder containing a survey, the Oregon Science Standards, prompts for reflections, and other supporting documentation. The actual scoring guides were added to the panelists' collection of documents at the conclusion of the first panel activity discussed below. Several panelists brought their personal notebook computers to the panel sessions. Digital copies of documents were made available to the panelists upon request.

After completing the brief survey, the first task the panels were assigned to complete was to review student work samples. The examples of student work samples used were gathered with the assistance of TOSAs working with CSE. These work samples represented the closest examples of student work samples that could be found, for which consent could be acquired, and that were of the

type of work samples the scoring guides were intended to score. The teachers were first instructed to review the student work samples and make a general assessment without the aid of a scoring guide of whether, in their opinion, the work sample demonstrated proficiency in SI or ED for the grade levels at which they taught. Once the panelists had completed their initial assessments they were offered the then current draft version of the scoring guide (V1.5) in a slightly modified format.

In the format provided by ODE, the scoring guide documents included both SI and ED scoring guides on the same page, side by side. For an example see Appendix C. The documents issued to the panelists for this event were modified to include only one scoring guide, SI or ED. See Appendix E. In lieu of a second scoring guide on the page, space was provided for the panelists to take notes as they evaluated the scoring guides and formulated the feedback they would suggest their panel recommend to ODE. The decision to limit the scoring guides to only SI or ED was made to help insure the panels focused only on the scoring guide their panel was designated to evaluate.

With the scoring guides in hand the panelists were asked to score the student work samples and to discuss the work samples, scores, and scoring guides. This placed the scoring guides in the context in which they were intended to be used. Before and after this activity, the panels were guided to reflect in writing on the scoring guides.

The following morning the panelists reconvened. The panels retained the same panelists and continued to focus on the scoring guide they had worked with

the previous day. The panels were instructed to come to a consensus within their panel for feedback, recommended changes, and rationales concerning the scoring guides they would like to offer ODE. The audio recordings from this session were captured and considered as data for this study.

### *Data Analysis*

The data collected in this study were analyzed through multiple methods ranging from document analysis to theme analysis. This study set out to answer two questions:

1. What types of feedback did the panels of experienced teachers offer ODE for the revision of the SI/ED scoring guides?
2. How did ODE utilize the feedback the teachers offered to revise the SI/ED Scoring Guides?

To answer these questions, the feedback documents and all the changes made in the several draft versions of the scoring guides needed to be identified, analyzed, and compared. The format in which the data was received shaped some of the analysis methods used to answer these questions.

*Formatting for Analysis.* All the scoring guides followed a similar format. Each scoring guide was comprised of four sub-sections. The titles for the four sub-sections of V1.5 of the high school scoring guides are provided in Table 6 as an example. Following the title was a brief description of the subsection.

**Table 6: Titles of Scoring Guide Subsections**

Science Inquiry	Forming a Question or Hypothesis	Designing an Investigation	Collecting and Presenting Data	Analyzing data, interpreting results, and communicating knowledge
Engineering Design	Forming a Question or Hypothesis	Generate possible solutions	Testing Solution(s) and Collecting Data	Analyzing data, interpreting results, and communicating findings.

All of the scoring guides in V1.5 had 6 scale degrees, or possible scores, indicating different levels of proficiency for each subsection within the scoring guide. Within each scale degree there were several sentences separated by bullet points which defined the required level of demonstrated understanding or ability for that score. Typically for any given scale degree, or score, there were three bullet points with a few exceptions in the elementary school scoring guides containing one or two bullet points

To facilitate analysis comparing the feedback documents to the scoring guide draft versions, these documents were migrated into a set of spreadsheets. Each scoring guide draft version and set of recommended changes from the teacher panel feedback documents were placed in a column of a spreadsheet with additional columns inserted as necessary to hold codes and notes for the researcher. Each block of text in the scoring guide was aligned in rows so that the blocks of text in every cell along that row were the from the same subsection,

the same scale degree, and the same bullet point of any scoring guide or feedback item in the spreadsheet. During the coding process there were 6 spreadsheets, one for each panel.

As additional draft versions of the scoring guides were released, the next revision of the scoring guide was placed in a new column of the appropriate spreadsheet. In the end each spreadsheet had 6 columns containing draft version V1.5, the feedback from the Teacher Evaluation Panels, and scoring guides V1.6 through V1.9. Intermingled among these columns containing the documents which would be analyzed, there were several additional columns to be used by the researcher for notes and codes. After the coding process described below was completed, the six spreadsheets were migrated again into a flat file data structure to enable further comparisons and analyses of the data for all of the scoring guides in aggregate. See Appendix H.

*Units of Analysis.* All panels chose to use the scoring guide document provided to them in their packet as a template to draft their recommendations and feedback on the scoring guides. The panels used the templates in different ways with variations in format and method. However, an artifact of this choice was that most of the feedback was formatted by the panels to match up with the scoring guides bullet for bullet. Each bullet point, description, and subsection title formed a unique block of text that could be tracked and compared between the scoring guides and the Teacher Evaluation Panels' feedback. Not all of the feedback was pre-aligned with the blocks of text which the feedback addressed. Some of the feedback was very general and did not pertain to any specific

block(s) of text. Some of the feedback did address specific block(s) of text. In the latter case, there was a column in the spreadsheets dedicated to additional notes from the Teacher Evaluation Panels. In the former case, the general feedback was left as a note at the bottom of the spreadsheets used for coding feedback.

The initial analysis was a comparative analysis that looked at the feedback and scoring guide draft versions V1.5 and V1.6. The changes made between V1.5 and V1.6 were identified using a function in Excel that compared one text string, or block of text, character by character and would flag a row if there was any difference between the two blocks of text.

Once changes were identified a series of Boolean comparisons were made comparing instances of change in the scoring guides to instances of feedback. This analysis produced results that identified and tabulated instances when there was feedback and a possibly correlative change to the scoring guides. However, it was noted early on that the content of some changes did not necessarily match some, or in many cases any part, of the content within the corresponding feedback. See example further below.

A unit of analysis with a finer resolution than the blocks of text mentioned above would be needed in order to answer the research questions. A method needed to be developed that could capture the content of the changes made and the content of the feedback in order to allow for meaningful comparisons. In addition to differences of content, within many blocks of text there were several distinct changes, or recommended changes that could be identified. Each

identified change or subject of recommended change became a distinct item that could be compared between the feedback and the scoring guide draft revisions. These distinct items became the units of analysis for this study.

Each distinct item was given its own row within the formatted spreadsheet. The text of the row was duplicated for as many distinct items as were identified. This enabled the researcher to keep track of each instance of change or feedback and maintain the context in which these changes or recommendations for change occurred. For example in the feedback from the elementary school SI panel regarding the description for the subsection "Analyzing and Interpreting Results" there were three distinct recommended changes, items, identified.

Draft V1.5:

Summarize, analyze and interpret data from investigations that address the identified question.

Feedback:

Summarize, analyze and interpret *patterns in* data from an investigation *or experiment* that address the identified question or hypothesis.

Draft V1.6:

Summarize, analyze and interpret data from an investigation that address the identified question or hypothesis.

The three distinct items in the feedback were the identified in italics above: the addition of "patterns in", the addition of "or experiment" and the addition of "or hypothesis". However, only one item of change between V1.5 and V1.6 was identified: "or hypothesis" -- marked with an underline in the above example. The row of the spreadsheet containing this block of text, the description for the



"Analyzing and Interpreting Results" subsection, was duplicated so that there were three rows with identical blocks of text for each row and a separate code to account each distinct item.

*Matching items.* One of the most important steps during analysis was to relate feedback items with change items in order to gather evidence that the feedback was used by ODE to make changes to the scoring guides. An example of a matching change was also included in the example above. Among the three items that were identified in the feedback, one of the three was identified as having a positive correlation with a change item, again noted with the underlined segment. These instances where the feedback and the change match in content are discussed as a 'match' in this study.

#### *Developing the Categorization Matrix*

After all the change items were identified and formatted for individual analysis, the feedback and draft versions of the scoring guides were analyzed again. This round of analysis focused on common themes to identify general types of changes and feedback. There were four general types of feedback and changes identified during this process. The themes that became apparent were categorized as Other, Structural, Evidence for Proficiency Score, and Performance Objective. Within these four categories three subcategories, or types, became apparent. To account for these subcategories the instrument being used to guide the coding process became a 2-dimensional Categorization Matrix. See Appendix A.

A coding scheme was developed based on this matrix. The main categories were identified by a letter: a, b, c, or d. The types within a category were identified by a number: 1, 2, or 3. Table 7 provides an overview for each code in the matrix.

**Table 7: Overview of Coding Categories and Subcategories.**

Category	Code	Subcategory	Description
Other	a1	Grammatical change	Grammatical change that did not affect the meaning of the statement.
	a2	Unclear	Category is unclear from text of feedback document.
	a3	Redefined	Item is redefined and no longer comparable with previous scoring guide.
Structural	b1	Number of bullets	Adds or deletes the number of bullets describing a score.
	b2	Scale degree	Decreases the number of scale degrees of the scoring system.
	b3	Order of bullets	Changed the order of the bullets within a score description.
Evidence for Proficiency Score	c1	Clarified	Clarified the degree of evidence needed for proficiency score.
	c2	Reduced degree of evidence	Omitted degree of evidence needed for proficiency score but did not change the performance objective.
	c3	Increased degree of evidence	Added degree of evidence needed for proficiency score but did not change the performance objective.
Performance Objective	d1	Clarified	Clarified the performance objective to be scored for proficiency.
	d2	Reduced requirement	Omitted requirement for performance objective to be scored for proficiency.
	d3	Increased requirement	Added requirement for performance objective to be scored for proficiency.

All the feedback and changes to the rubrics were coded using the Categorization Matrix. In addition to coding the items, a short note describing the coded item was added to the adjacent column to aid the researcher as further comparative analysis was performed.

Once the coding process was complete, the whole data set was audited to confirm a uniform application of the coding criteria was applied. The next step was to perform a frequency analysis for each coded item in the data set. Each instance of a code was counted and cross tabulated to construct an image of what types of feedback the different panels offered ODE and a better picture of the changes that took place in the scoring guides.

*Total versus uniquely coded items.* There were two ways to count the feedback or change items: 1) to count all of coded items including repeated items or 2) to count only the instances of uniquely coded items. Unless otherwise stated all the frequency data analysis was computed using uniquely coded items.

A uniquely coded item was the first time an item was coded in a given context. The most common example of shared context for a repeated item to occur was when a coded item had the same content as an item previously coded in that subsection of the scoring guide. However, if an item was repeated in a different context, then it was still coded as a unique item.

An exception to this rule was with 'Structural' items in which case the context for the item was the whole document. For example, a 'b2' type feedback to omit the 5th scale degree from the scoring guide was offered. Since the

feedback item was universal for the whole document, that is, the 5th scale degree was to be removed from all four subsections of the document, there was a total of four 'b2' codes for this set of items. However, the feedback items were counted such that there was a uniquely coded item only for the first subsection in which the scale degree was to be omitted. The three remaining coded items with the same content were coded as repeated items since the context for the feedback items was the same. The same process was used when coding and counting structural change items in the scoring guides.

The next step in the process of analysis was to compare the codes between the lines of the feedback and V1.6 of the scoring guides. When the codes matched, the text blocks were flagged for follow up analysis to confirm whether the content of the change reflected the content of the feedback. If the content of the feedback and the change matched exactly the flag marking the match was left in place. If the content of the change only partially matched the content of the feedback, the flag was modified to indicate a partial match. If the content of the change did not match the content of the feedback, the flag was deleted meaning the feedback code and the change code matched coincidentally and there was no evidence for that feedback item affected the revision of the scoring guides.

The last step in analyzing the scoring guide documents was to code changes made to the following drafts of the scoring guides, V1.7 through V1.8. These changes were compared to the feedback and to V1.5 and V1.6 with an

eye to find additional themes that might identify the types of changes made to the scoring guides and generally how the scoring guides developed.

The final step of analysis was to listen to the audio recordings of the Teacher Evaluation Panels as they wrote their feedback documents. These discussions were analyzed for themes that might develop as well as gain better insight into the processes and rationales behind the feedback the Teacher Evaluation Panels developed.

## Results of Analysis

### *Overview*

The first question this study answered was what type of feedback the Teacher Evaluation Panels would offer ODE. The results below begin with a brief discussion of the processes and attitudes observed among the Teacher Evaluation Panels as they developed their feedback documents. This is followed by a closer inspection of the codes used to identify the types of feedback, as well as changes made, with examples to illustrate the types of feedback that were offered to ODE. Next the frequency of the identified types of feedback will be discussed followed by a similar discussion of the changes made to the scoring guides. Results of the comparative analysis between the feedback and the changes observed in the feedback answered the second research question which asked how the feedback would be used by ODE to further develop the scoring guides. This will be followed by data showing how the scoring guides developed after the initial feedback was accepted. Finally there will be some results from the analysis of the audio recordings of the teachers' discussions in the panels.

### *Audio Recordings of Teacher Evaluation Panels*

The audio data recorded during the session in which the Teacher Evaluation Panels developed the feedback documents later submitted to ODE provided some useful data. Upon listening to these recordings with the intent of identifying themes that would be used to characterize the types of feedback the panels offered, there is was very little new information gained. The bulk of the recorded

discussions contained the same general raw data as was captured in the feedback documents which they were writing. That is, for the most part the data that was captured in the audio recordings was the same data as was captured in the feedback documents.

There were a few cases in which discussions provided some additional insight into the teachers' thinking processes about some feedback items or their attitudes about the scoring guides. The original hope for these recordings was to capture more data about the feedback items themselves. However, in the end, the main value gained in the audio recordings was that it captured the actual processes the teachers used as they evaluated the scoring guides.

*The Evaluation Processes.* Though there were slight variations in the approaches taken by the different Teacher Evaluation Panels as they drafted their feedback documents, in general the panels worked with draft V1.5 in a methodical page by page manner. Some panels began discussing the different subsections at the proficiency score level -- 4, others began at the top score -- 6. During the feedback development session, all the panels discussed each block of text in the scoring guides. Sometimes the discussions were very short, such as when none of the teachers had any issues with a given block of text. Sometimes the discussions would carry on into collegial debates as the teachers worked out what they thought draft V1.5 was saying and what they wanted the scoring guide to say. Frequently these debates would engage the content of the scoring guides at a very nuanced level, working out the meaning both in terms of what the evidence the students would be expected to include in the work sample as

well as how the teachers would use the scoring guide to differentiate between one score versus another.

One such debate occurred within the B2-ED panel, discussing their feedback to highlight specific text in the scoring guides, their debate centered on whether to highlight the logical connectors or the key concepts linked by the connectors. Another discussion centered around the need for feedback regarding the wording in the high school SI scoring guide that tried to differentiate between "not clearly defined" and "not existing". The assertion by one panelist in the HS-SI panel was that it was hard to distinguish between these descriptors when looking at student work. Another debate present among the middle school and high school ED panels was the requirement that the 'problem' the students were to attempt to solve must be a "practical problem". See the first subsection title in Table 2: Scoring Guide Draft Revisions. The meaning and the extent of limitations defining practical versus non-practical problems was itself problematic in the view of these panels. The debates, or discussions, of differing opinions about the scoring guides or the feedback to be offered were respectful and were resolved with the apparent agreement of all the members of the panels. The debates in these cases captured the care and detailed nature in which the feedback was formulated.

Not all of the feedback ideas offered by the panelists were adopted by their panels to be included in the feedback documents. For example, the middle school SI panel had a discussion about the reordering the bullet points as well as



labeling the bullet points a, b, c. After a brief discussion, neither of these suggestions were carried forward into the feedback document submitted to ODE.

The expressed attitudes of the teachers in the evaluation panels during the development of the feedback documents were generally positive both toward the scoring guides and toward the processes of working together with their panels. One panelist commented that "going over the scoring guide line by line was exactly what I wanted to do." Another panelist said that the meeting of these panels was the best professional development event in which that panelist had ever been a participant, even though professional development was not the designed intent for convening the panels.

However, there was also a theme that most panelists were aware of the limitations of these sessions in terms of time that was available and what they would be able to achieve within these panels. For example, one panelist stated "We could change the whole (scoring guide), but we need to change the things that are the most problematic." There was also unanimous consensus from all panels that they would like to see a vertical articulation of the scoring guides, working out the expectations and transitions between the multiple grade levels at which the scoring guides would be used.

One panel held a unique view of the scope, or degree of freedom, they had with respect to the feedback they would offer. The middle school ED panel, while recognizing the time limitations, decided to provide the most feedback they could, working from the notion that "this is a new scoring guide" this panel took "the opportunity to start fresh." This degree of freedom expressed by this panel

was clearly contrasted with the final feedback of the high school ED panel. This latter panel had verbally expressed a desire to include a flow chart version of the scoring guides in their feedback document to further emphasize the reiterative process of engineering, but ultimately did not include that recommendation with an expressed doubt that the recommendation would be accepted by ODE.

Another type of data offered in the audio recordings that were not captured in the feedback documents were feedback items that were discussed, but were not included with the feedback documents that were submitted to ODE. There was a brief mention of an additional document that the high school SI panel was working with but which did not get included with the panel's feedback document. There were a couple instances where panels had lengthy discussions about a particular item of interest which sounded as if it was intended to be included in the feedback documents, but in the end these items were not present in the feedback that was submitted to ODE. However, the focus of this study is on the feedback that was submitted to ODE and how ODE used that feedback, so the hypothetical effects of the lost feedback items were not considered further beyond noting that the feedback submitted to ODE may have been incomplete.

A final type of data found in the audio recordings was clarification of the meaning of some items of feedback. There were a few instances in the feedback documents where the feedback was unclear or difficult to interpret based on the text and context. In these rare cases, the audio recordings occasionally offered some assistance in clarifying what was meant or intended by an item of feedback. While these data points were helpful in discovering the meaning of

certain items of feedback, this study is, again, focused on the feedback that was submitted. Feedback items that were unclear to the researcher were coded as 'a2' and were included in all stages of analysis including the comparative analysis seeking connections between items of feedback and items of change in the revised scoring guides.

### *Examples of Coded Items*

The first question this study set out to answer was what types of feedback the Teacher Evaluation Panels would offer ODE. To answer this question a Categorization Matrix was developed through a process of document theme analysis. Within the Categorization Matrix there were 12 codes described, one for each type of feedback or change item identified in this study. Eleven of these codes were identified from feedback items. One code from the matrix was developed for a small set of changes that occurred in draft V1.8 of the scoring guides but was not present in the feedback from the Teacher Evaluation Panels.

Since this study was first and foremost concerned with the feedback the Teacher Evaluation Panels submitted to ODE, the examples below are from the feedback documents with the one exception mentioned above. The coded examples are typical of items assigned that code, and in the case of some codes, are exhaustive.

It is important to note here that there were many cases in which feedback was offered or changes made to the scoring guides, there were more than one item of feedback or change identified. This is true for many of the examples used below. It was quite common that there would be more than one coded item

identified in a block of text. In the interest of brevity, the examples below only discuss the coded item relevant to that section. An example of the breakdown of a single block of text into multiple coded items follows the itemized examples of coded items.

*Other - 'a1'*. The feedback coded 'a1' -- "Other - Grammatical" were grammatical or language changes which did not change the meaning of the item. Of all items coded 'a1', most of the items were changes in the scoring guides. There were only a few items coded 'a1' in the feedback offered by the Teacher Evaluation Panels. The example below is typical of an 'a1' feedback item:

Panel	Draft V1.5	Code	Researcher Note	Feedback
B3-SI	Provides observations and/or scientific principles that relate to the question or hypothesis.	a1	voice	Background observations and/or scientific principles relate to the question or hypothesis.

When the text block of the scoring guide was compared to the feedback, there was a substantial change. However, the overall meaning of the text block did not change beyond the voice of the text. In this example, the meaning of the text blocks with respect to what the students work sample needed to show did not change. The requirements remained the same, therefore the item was coded 'a1'. There was a total of 41 unique 'a1' type items identified in this study, most of which were found in the scoring guide draft revisions.

*Other - 'a2'*. There were only 3 items in the study coded as 'a2' , "Other - Unclear". All of these items were feedback items and in all cases it was unclear

to the researcher as to what the feedback as expressed in the feedback document was addressing or what change the feedback was recommending.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B2-ED	Thoroughly records the results from testing the solution and identify unexpected outcomes.	a2	strange symbol (?)	(?) = ... in graphs and writing...

Additional clues about the intent of these items were found in the audio recordings. For example, in the case of the B2-ED feedback item above, it was not intelligible as to which block of text the feedback was intended to be attached to or what the symbol meant until listening to the section of the audio recording in which they were discussing the feedback item. The strange symbol was finally deciphered as an 'insert here' symbol only with the aid of the recording. The coding of all feedback items was done purely on the basis of the textual evidence that could be gleaned from the feedback documents that were presented to ODE. Even if the feedback item could be interpreted after considering the additional data captured in the audio recordings, if the item could not be interpreted from the document and reasonably interpreted by the ODE Science Content and Assessment Panels, the items were left with the original code 'a2'. In the final analysis all 'a2' items remained identified as 'unclear'.

*Other - 'a3'.* There were 11 'a3' items identified in this study. All of these items were feedback items from the middle school ED panel. Items coded as

'a3', "Other - Redefined" posed the most significant challenge when coding the feedback and scoring guides. Items coded a3 were feedback items that generally addressed the performance objective and could have been coded as a 'd' type items. However, these feedback items also represented a more extreme departure from other feedback items more typical of the 'd' type. The difference between these types was substantial enough that they needed to be identified with a code that distinguished these feedback items from other coded feedback items.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B3-ED	Describes a variety of possible solutions that are distinctly different.	a3	propose sol'n --> design process	Creates a decision tool to analyze all reasonable solutions in terms of the criteria and constraints.

In the example above, the text block from V1.5 outlines the requirement that the student's work sample describes solutions to the problem. The feedback item recommended that the requirement change to a process in which the student "creates a decision tool" that could be used to analyze the solution(s). This item could not be classified in the same category as the feedback more typical 'Performance Objective', 'd' type feedback. See examples for category 'd' items below. Therefore, items such as these were coded as an 'a3'.

*Structural - 'b1'*. Generally items coded in the 'Structural' category were items that changed the shape or organization of the document. There were 10 unique items coded as 'b1', 'Structural - Number of Bullets', all of which were

feedback items, and all but two of which were from the middle school ED panel. Typically, the recommended change was simply to omit or add a new bullet point, or piece of evidence required to demonstrate a level of proficiency. The examples below were taken from different sections of the B3-ED scoring guide analysis. The first example shows a case where it was recommended that a block of text be removed from the scoring guide. The second example shows a recommendation to add a bullet that split the content of a text block in V1.5 into two separate text blocks.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B3-ED	Constructs a solution that adequately addresses the criteria and constraints and is appropriate for testing.	b1	delete bullet	
B3-ED		b1	add bullet, content from bullet 2	The design fits with both the criteria and constraints.

In many cases in which there was an added bullet point, the text block had multiple codes, one of which was as a structural item and then other codes for the content of any additional changes within that block of text.

*Structural - 'b2'*. The 17 unique items coded as 'b2', or 'Structural - Scale Degree' were items where the scale degree was removed from scoring guide. Draft V1.5 of the scoring guides had six scale degrees, 1 through 6, each with its own set of requirements detailing the degree of evidence required for the student

to earn that score. The feedback items and the change items that were coded 'b2' omitted an individually detailed scale degree from the scoring guides. The example below is the feedback item from the high school ED panel to remove the score of 5 from the scoring guide.

Panel	Draft V1.5	Code	Researcher Note	Feedback
HS-ED	Describes in detail a practical problem to be solved through the process of engineering design, clearly tying constraints and science principles to the problem.	b2	omit scale degree	

Only two panels offered feedback of this type. The feedback from these two panels was unanimous about removing the 5th and the 1st scale degrees from the scoring guides. However, the panels differed in that one panel also recommended universally removing the 2nd scale degree as well. The feedback from the middle school ED panel also wanted to remove a third scale degree leaving the scoring guides with 3 scale degrees -- Emerging, Proficient, and Exceeds. The scoring guides were changed into a hybrid 4-scale degree scoring guide, which retained the 4th scale degree as the mark of proficiency. See Appendix D. The detailed requirements for the 5th and 1st scale degrees were removed from the scoring guides, however the top and bottom scale degrees of the scoring guide read '6/5' and '2/1' respectively. This allowed the scorer to continue to differentiate between degrees of exceeding proficiency, 5 or 6, or the



degree of emergent demonstrated ability, 2 or 1, based on the evidence presented in the work sample.

*Structural - 'b3'*. Only one item was coded as 'b3', 'Structural - Order of Bullets'. It was from draft V1.8. There were multiple changes that occurred within this block of text, as was frequently the case with many blocks of text which contained change items or feedback items. The example below only includes the unique structural code as the example of this type of change. This change, switching the order of the bullet points, was also an example of a set of repeated codes. For example, the switch of order of the bullet points was carried out for all scale degrees in that subsection of the scoring guide. Only the first instance was coded as a unique instance. All other instances of the switched order of bullets were coded as repeats. The example below shows how the actual text changed and how the alignment of the codes and notes followed the change for further downstream analysis with newer scoring guides.

Panel	Draft V1.7	Code	Researcher Note	Draft V1.8
B2-SI	Uses relevant scientific knowledge and principles from multiple sources to independently frame an investigation.	b3	change order of bullets 1 & 2	Forms a testable question or forms a hypothesis that clearly guides the design of a scientific investigation.
	Generates a testable question or formulates a hypothesis that clearly guides the design of a scientific investigation.	d3	add 'observations'	Uses specific observations and relevant scientific principles from multiple sources to independently frame an investigation.

*Evidence for Proficiency - 'c1'.* Items coded in this category were items for which the requirement in the block of text was changed to either clarify, decrease, or increase the evidence required to earn the score represented. Items coded as 'c1' were a clarification for that block of text. For example, the item below was coded 'c1' because it made the requirement of 'data quality' described in V1.5 clearer. However, this recommended change only affected the block of text, the bullet point, requirements for a score of 6, as "data quality" was not a requirement for scores below a 6. There was a total of 58 unique 'c1' type items in this study.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B3-SI	Rigorously follows the specified procedure, monitors data quality and utilizes the best available tools and techniques.	c1	clarifies 'data' quality	Data collected is consistent with the procedures and is precise, accurate, sufficient.

*Evidence for Proficiency - 'c2' and 'c3'.* Items coded 'c2' or 'c3' respectively omitted or added a requirement that affected the score in which the item was coded but did not change the overall performance objective for the subsection. There were 63 and 78 unique items for 'c2' and 'c3' respectively. The examples below either omitted a requirement or added a requirement that was unique to the proficiency score the block of text described.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B3-SI	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c2	omits comprehensive	Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.
HS-SI	Provides comprehensive background science knowledge and observations to establish a detailed context for this investigation.	c3	add with citations or well documented	Comprehensive quality- defined with rubric that describes acceptable as "with citations", or well documented personal/human experience

*Performance Objective - 'd1'*. The 'Performance Objective' category was for feedback or change items that addressed the overall objective the performance assessment would measure students' demonstrated proficiency through the work sample. The 58 unique items coded with a 'd1' were items in which a clarification of the performance objective on the whole was made rather than just a clarification for a specific score requirement. For example, below an item was coded 'd1' in which draft V1.5 referred to "tools". The change recommended by the Teacher Evaluation Panels was to change "tools" to "resources/materials", clarifying what "tools" the students would be expected to use in their work sample. This recommended change was made throughout the subsection for each instance where "tools" was used in this way. Each subsequent item addressing this clarification after the first instance was coded as

a repeated item within the document. When the same item was identified in the following draft of the scoring guides, the first instance in the new document, draft V1.6, was again coded as a unique item with subsequent items coded as repeats of the first instance.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B3-SI	Designs a scientific investigation that uses appropriate tools and techniques to collect data relevant to the question or hypothesis.	d1	tools' --> 'resources/materials'	Designs a scientific investigation that uses appropriate resources/materials and techniques to collect data relevant to the question or hypothesis.

*Performance Objective - 'd2' and 'd3'.* Similar to the 'c' category, items coded with a 'd2' or 'd3' respectively omitted or added a requirement. However, for the 'd' category, the requirement affected the performance objective on the whole rather than simply the evidence required to achieve a particular score. The examples below either omitted or added a requirement that changed the performance objective of which the students were expected to demonstrate proficiency. There were 56 unique 'd2' items and 41 unique 'd3' items identified in this study.

Panel	Draft V1.5	Code	Researcher Note	Feedback
B2-SI	Design a scientific investigation to answer questions or test hypotheses using appropriate tools and	d2	'questions' --> 'question'	Design a scientific investigation to answer a question or test a hypothesis using appropriate tools and

procedures.

procedures.

HS-SI	Thoroughly identifies relevant variables and defines a systematic investigative process that is clearly defined and adaptable if necessary.	d3	add 'controls and monitors'	Thoroughly identifies relevant variables and defines a systematic investigative process that is clearly defined and adaptable if necessary.
-------	---	----	-----------------------------	---

*Text Block with Multiple Codes.* Many, if not most, of the text blocks which had items of feedback or changes identified had more than one item identified for the same block of text. The example below was typical. Though there was only one block of text in either the draft V1.5 or the feedback showing the recommended changes, there were 3 distinct ideas represented in the feedback from this panel. In order to capture all the types of feedback, and changes observed, it was necessary to split the instance of a changed block of text into separate items to capture the different ideas represented. Mechanically, this was done by duplicating the row within the coding data structure, see Appendix H. When the data was migrated to the analysis data structure, see Appendix I, the rows were again duplicated. This allowed for these items to be counted individually as well as cross referenced individually with change items coded in downstream scoring guide draft versions.

Panel	Draft V1.5	Code	Researcher Note	Feedback
	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c3	adds appropriate	Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.
B3-SI	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c3	clarifies bg knowledge	Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.
	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c2	omits comprehensive	Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.

### *Results of Feedback Analysis*

Following the process of coding the feedback items using the Categorization Matrix was a process of counting instances of feedback items and comparing these results to identify patterns in the data. Table 8 provides an

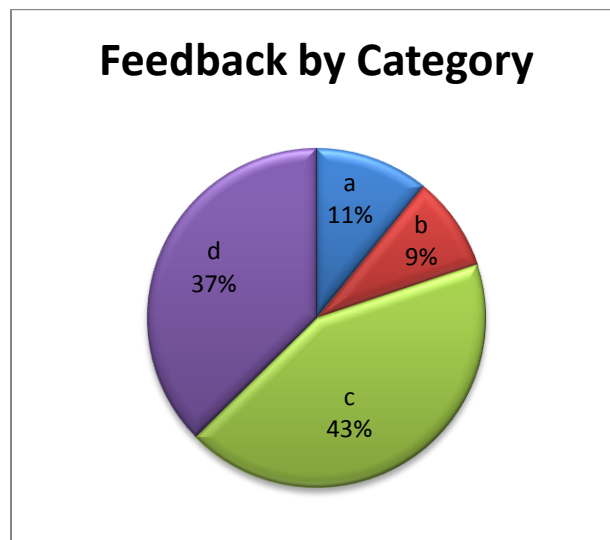
overview of the frequency of coded feedback items. The instances of each individual code were tabulated along with aggregated sums for each category, both for the total number of instances -- including repeated codes and uniquely coded items. Distinguishing unique versus repeated items had the greatest impact in the Structural and the Performance Objective categories 'b' and 'd' respectively. Due to the nature of these categories there was a higher likelihood for repeated items as content of these types of were often carried forward to each level of the scoring guide. Unless otherwise noted, results in the graphs and tables below were analyzed using uniquely coded items.

**Table 8: Number of unique feedback items by category and type.**

Category	Code	Total	Unique
Other	a1	32	5
	a2		2
	a3		25
Structural	b1	30	10
	b2		20
	b3		0
Evidence for Proficiency Score	c1	87	26
	c2		33
	c3		28
Performance Objective	d1	145	63
	d2		55
	d3		27

*Types of Feedback offered by the Panels.* When the feedback the Teacher Evaluation Panels offered to ODE was looked at in terms of proportionality, 80% of that feedback pertained to the fundamental functionality of

the scoring guides: distinguishing between scoring guide levels, category 'c', or defining the performance objective, category 'd'. See Figure 2.



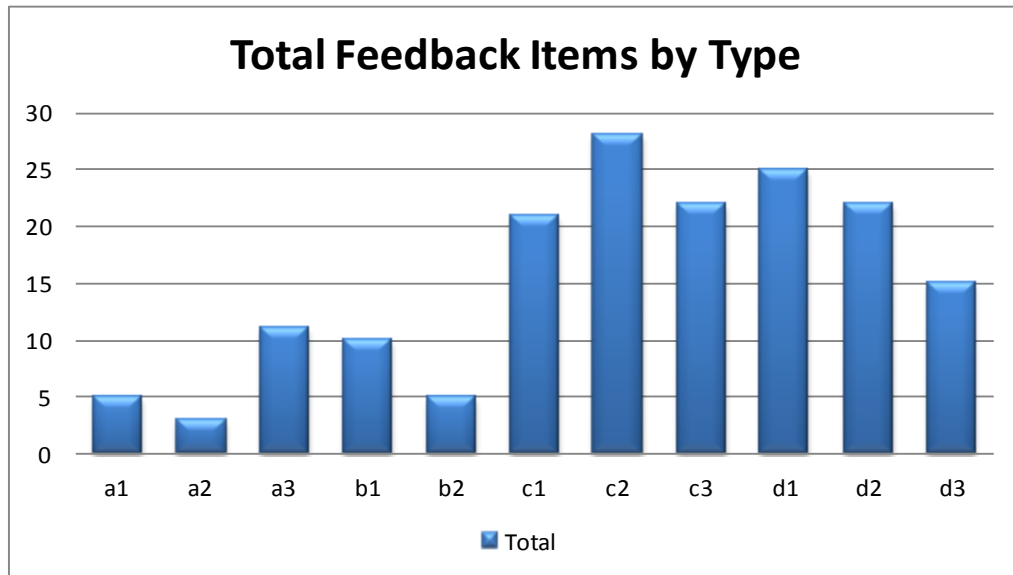
**Figure 2: Feedback items by general category**

This result was corroborated by the audio recordings of the Teacher Evaluation Panels' as they drafted the feedback they would offer ODE. Discussions of grammatical issues, type 'a1' and structural items, category 'b', were rare compared to discussions concerning the language of the scoring guide that addressed the performance requirements -- categories 'c' - Evidence and 'd' - Performance Objective, specified in the scoring guides.

A more refined look at the frequency data in Figure 3 shows the overall distribution of uniquely coded feedback items analyzed by individual codes. As shown in Figure 2, most of the feedback items were coded in categories 'c' - Evidence and 'd' - Performance Objective. Of these 'd3' - add requirement for performance objective had notably the fewer instances than other codes in the 'c' and 'd' categories, however, there were substantially still more instances of 'd3'

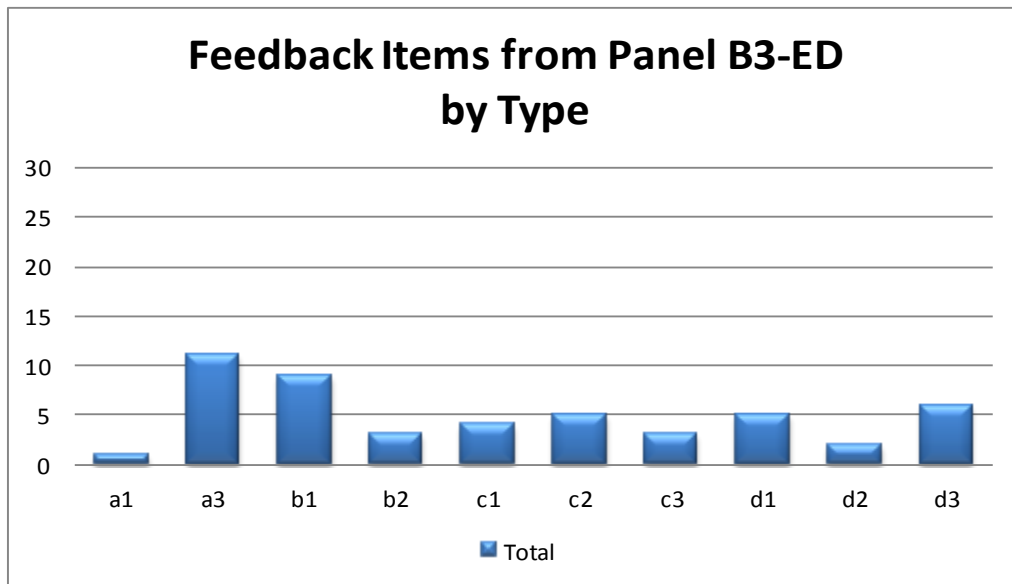


than any coded item in the 'a' - Other or 'b' - Structural categories. Except for item type 'b3' - change order of bullet points, there were feedback items of all types within the Category Matrix.



**Figure 3: Distribution of coded feedback items by type.**

The middle school Engineering Design panel (B3-ED) was the only panel to offer feedback that was coded 'a3', Other - Redefined or 'b1', Structural - Number of Bullets. Feedback of these types represented a large proportion of the total feedback offered by the middle school ED panel. See Figure 4.



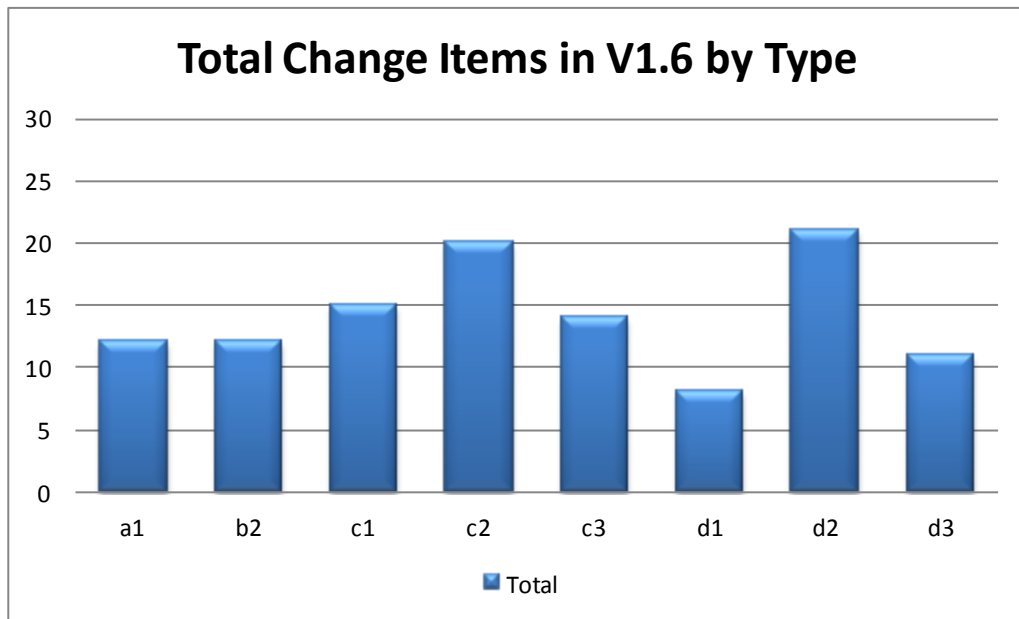
**Figure 4: Distribution of coded feedback items from Panel B3-ED by type.**

*Changes resulting in V1.6.* Based on the interview with the principle staff member of the ODE Science and Assessment Panels, it was known that the ODE Science Content and Assessment Panels made modifications to the scoring guides after they had received the feedback from the Teacher Evaluation Panels, and the feedback from the Teacher Evaluation Panels was utilized when making modifications to the scoring guides. The changes made during this revision cycle were released in draft V1.6. The same method and categorization matrix was used to analyze the changes to the scoring guides as was used to analyze the feedback offered by the Teacher Evaluation Panels. Like the feedback analysis, distinguishing between unique and repeated items had the most impact when looking at change items of the type 'b2' - Scale Degree and category 'd' - Performance Objective.

**Table 9: Coded changes in scoring guide drafts V1.6 by type.**

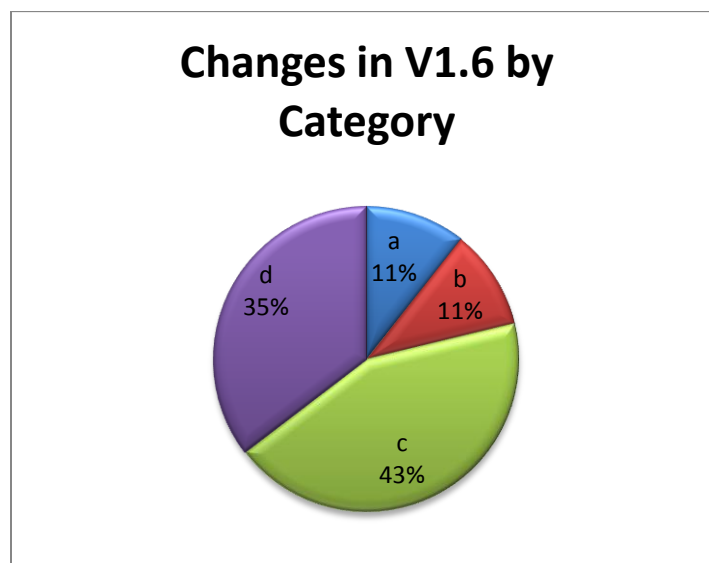
Category	Code	total		unique	
Other	a1	12	12	12	12
	a2		0		0
	a3		0		0
Structural	b1	51	0	12	0
	b2		51		12
	b3		0		0
Evidence for Proficiency Score	c1	49	15	49	15
	c2		20		20
	c3		14		14
Performance Objective	d1	71	14	40	8
	d2		37		21
	d3		20		11

Figure 5 shows a detailed look at the frequency data for the overall distribution of uniquely coded change items in draft V1.6 by type. There were fewer types of changes observed in draft V1.6 compared to the feedback. The



**Figure 5: Distribution of coded changes in V1.6 by type.**

most infrequent change observed were clarifications to the performance objective, type 'd1'. However, as shown in Figure 6, change items were predominantly in the categories 'c' - Evidence and 'd' - Performance Objective. The proportionality of change items by category compares very closely to the proportionality of feedback items. See Figure 2. Roughly 20% of the changes were type 'a' and 'b' changes and roughly 80% were of either 'c' or 'd' type changes.



**Figure 6: Proportionality of change items by category.**

*Comparing Feedback to Changes in V1.6.* A closer comparison of the quantity and the content of the feedback and change items is considered, the similarity between the proportions of feedback items and change items in V1.6 becomes less meaningful. This similarity between Figure 2 and Figure 6 may only indicate that the Teacher Evaluation Panels and the ODE Science Content

and Assessment Panels were concerned with similar categories of issues within the scoring guides.

In general, there were far fewer change items identified in V1.6 compared to feedback items offered by the Teacher Evaluation Panels. See Figure 7 and Figure 8.

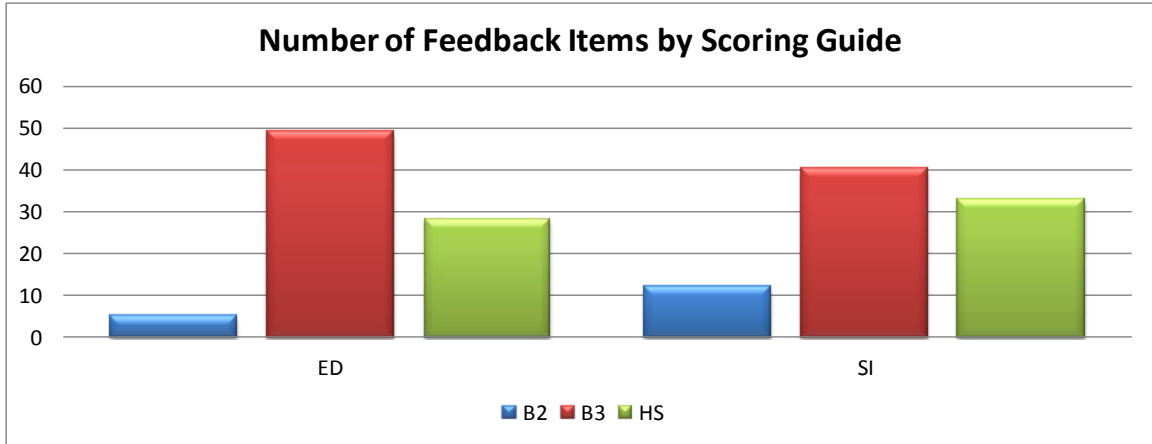


Figure 7: Total number of unique feedback items by scoring guide.

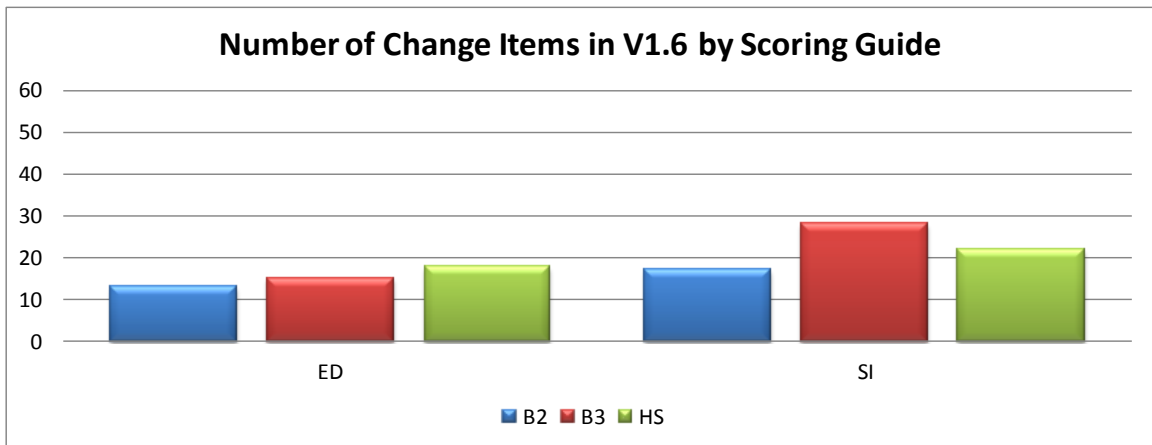


Figure 8: Total number of unique change items in V1.6 by Scoring Guide.

The number of coded feedback items varied for different teacher panels. See Figure 7. The elementary panels in both the SI and ED panels offered fewer

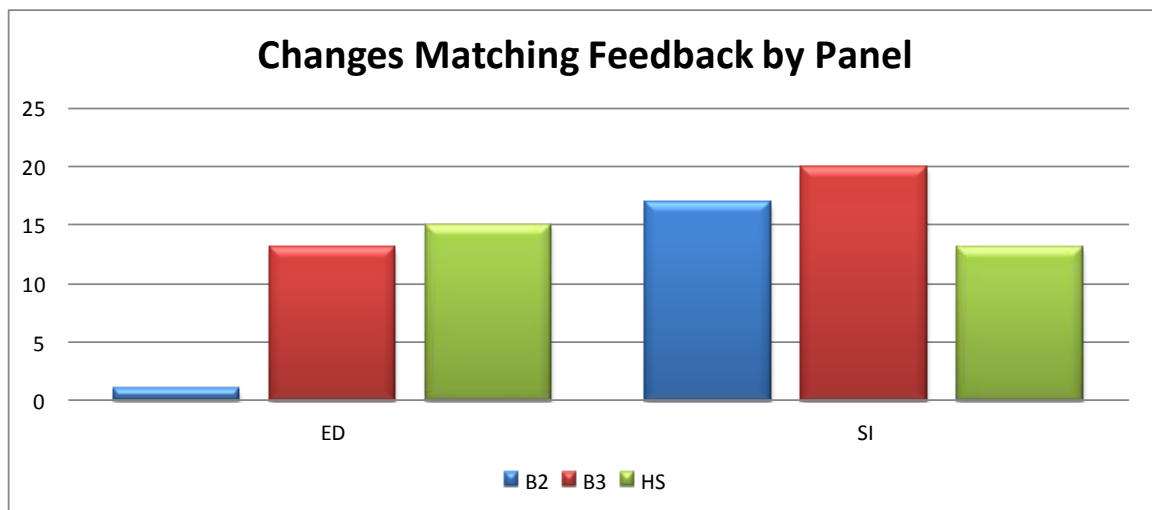
items of feedback than the other panels, with the elementary ED panel offering the fewest, 5, uniquely coded items. The middle school panels offered the most uniquely coded items of all the grade levels and there was close parity between the SI and ED panels teaching at the middle school level. The high school panels offered fewer unique feedback items than the middle school panels and more than the elementary school panels. The high school panels were also relatively even between SI and ED.

Further there were notable variations in the quantities of feedback items and the change items for the panels at different grade levels. For example, B2-ED and B2-SI both had fewer items of feedback than there were changes in draft V1.6. The middle school panels offered considerably more feedback than either the elementary school or the high school panels, yet the number of changes made to the scoring guides in draft V1.6 was relatively even across grade levels.

However, despite these differences, there was strong evidence in the interview with the ODE member that the ODE Science Content and Assessment Panels used the Teacher Evaluation Panels' feedback when drafting V1.6. When asked how the feedback was utilized the ODE member responded "in fact that was the feedback we used to proceed to the (V)1.6 work. I believe that out of that work, one of the largest changes that resulted from the Portland State meetings we had was the movement to a 4 level scoring guide that included the flexibility of being a 6 point scale." Identifying the changes that had the same content, that matched, the feedback was the next step in the analysis.

*Changes Matching Feedback.* Of the 166 unique feedback items offered by the Teacher Evaluation Panels, there were 42 unique change items in draft V1.6 that matched feedback items. If matches are interpreted as the adoption of feedback by ODE, then this is approximately a 25% adoption rate. There were 113 unique changes identified in V1.6. Given the same interpretation of adoption, then the ratio of change items to change items matching feedback items, 113:42 yields an interpretation that approximately 37% of the changes were a result of teachers' feedback.

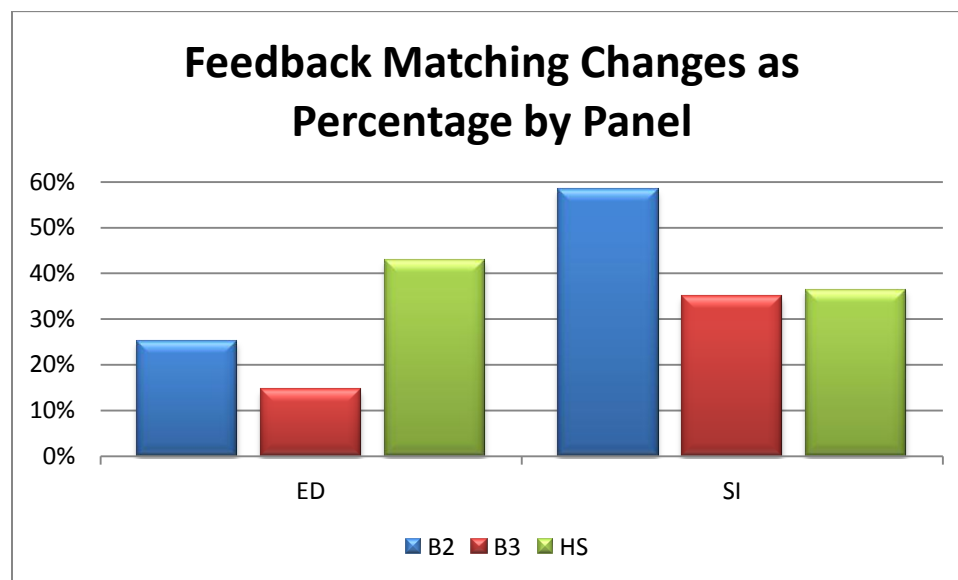
The frequency of matching changes varied significantly from panel to panel. See Figure 9. The elementary ED panel had only one feedback item that matched a change item in draft V1.6.



**Figure 9: Number of changes in V1.6 matching feedback.**

However, the frequency of teacher feedback matching change items in V1.6 as a percentage provides more insight into comparisons of how the different panels possibly affected the scoring guides. See Figure 10. These data show

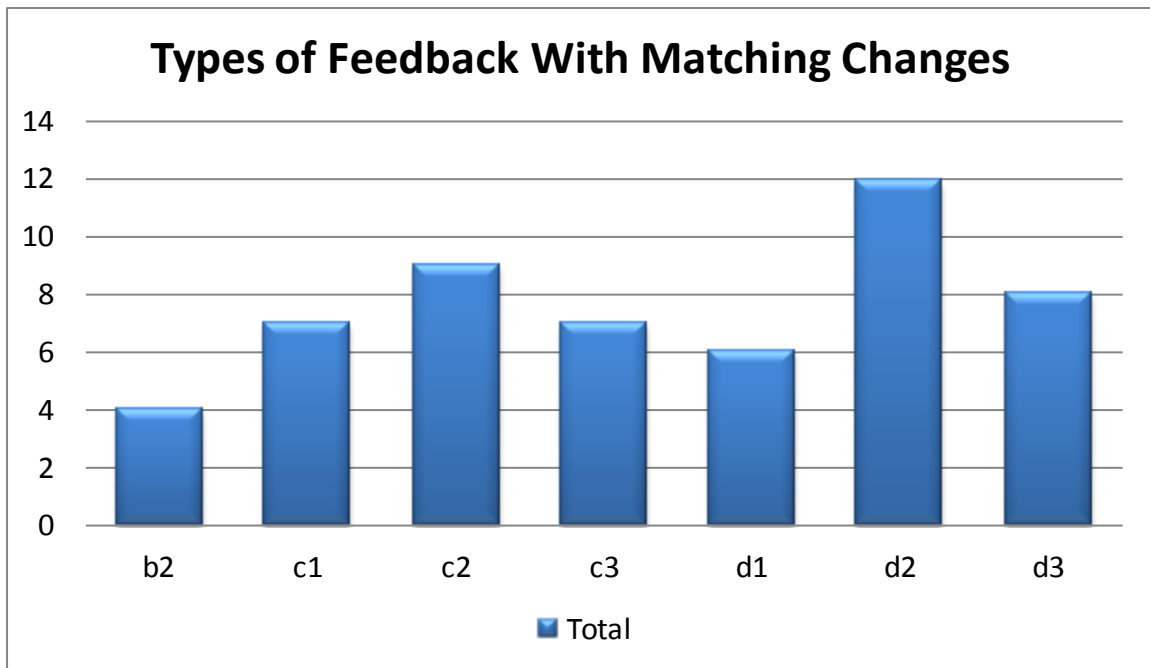
the correspondence rate of feedback matching changes as 25% or greater for all panels except for the middle school ED panel. The relatively low correspondence rate for the middle school ED panel may be explained by the fact that this panel offered the most feedback items, see Figure 7, and a large proportion, approximately 41%, of the middle school ED panel's feedback items were of the more extreme feedback types, 'a3' and 'b1', see Figure 3. No change items were identified in any scoring guide that matched these types of feedback.



**Figure 10: Feedback items matching change items as a percentage by panel.**

The frequency of feedback items with matching changes in draft V1.6 showed a wide range of frequency and variation by type amongst the different panels. See Figure 11. A wide variety of coded items from the 'c' - Evidence and 'd' - Performance Objective categories had matching changes, however, the only feedback items that also had matching changes from the 'a' or 'b' categories were of the 'b2' type, Structural - Scale Degree.



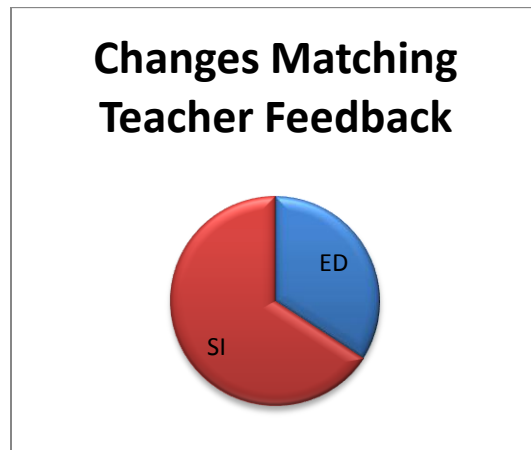


**Figure 11: Types of feedback with matching changes in V1.6.**

The number of changes coded 'c2' and 'd2' were more frequent than any other change item. The highest incidence of these codes occurred for the SI scoring guides. In the case of items coded 'c2', half of the items were for the middle school scoring guide. In the case of items coded 'd2', nearly half of the items were from the elementary school scoring guide. Both of these codes denoted a reduction of requirements. In the case of middle school, it was a reduction of evidence for specific scores within the scoring guide. In the case of elementary school, it was a reduction of the requirements the students would be expected to perform.

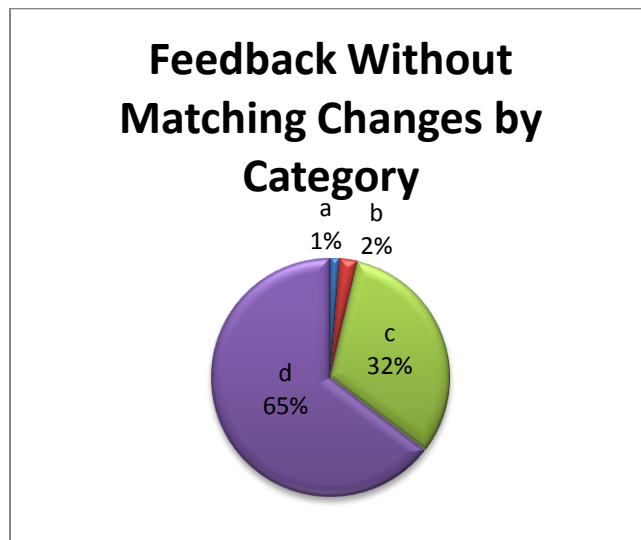
Though the number of unique feedback items between the SI and ED scoring guides were roughly equal, there was a clear difference in the proportionality of matching changes between the two types of scoring guides.

The changes in the SI scoring guides matched the feedback much more frequently than the ED scoring guides. See Figure 12.



**Figure 12: Proportion of change items matching feedback by Science Inquiry or Engineering Design.**

The types of feedback items that did not have matching changes in draft V1.6 were predominantly from the categories 'c' - Evidence for Proficiency and 'd' -- Performance Objective. See Figure 13. There were relatively few feedback items in the 'a' - Other and 'b' - Structural categories. Feedback in categories



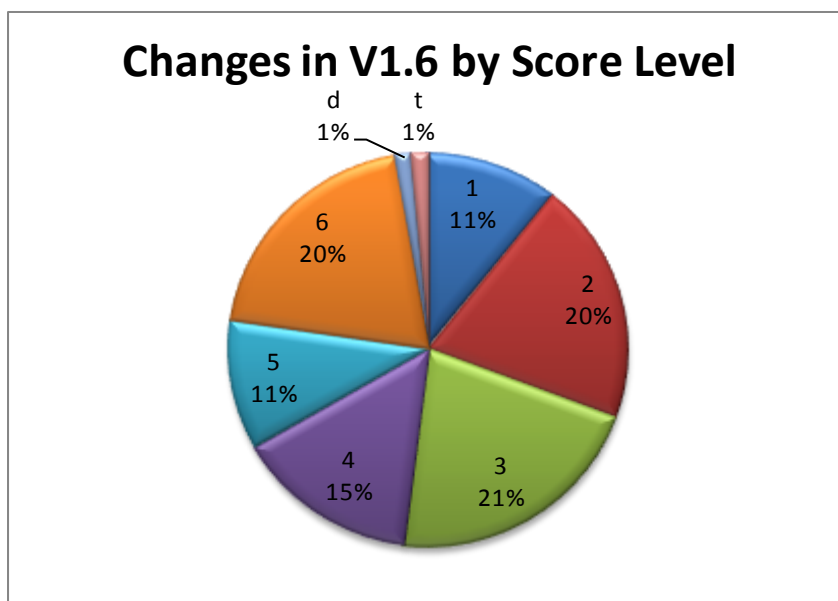
**Figure 13: Categories of feedback without a matching change in draft V1.6.**

'a' and 'b' were proportionately more often matched with changes in draft V1.6. These two categories were roughly equal in terms of feedback offered, see Figure 2. However, there was a clear disproportionality when looking at items with matching changes.

Finally, when comparing the score levels at which changes occurred, there was a marked similarity in all draft versions of the scoring guides and the feedback from the Teacher Evaluation Panels. See Figure 14 and Figure 15. Depending on the document - set of feedback or draft versions, the total number of coded items varies. However, a large proportion of coded items were located at score levels 3 to 6. Of the feedback items located at score level 5, half of those were 'b2' - Change scale degree items. Almost all feedback items at score level 1 were 'b2' items. All of the change items at score levels 1 and 5 were 'b2' changes. Very few items were identified in the subsection titles or descriptions.



Figure 14: Proportions of feedback items by score level.



**Figure 15: Proportion of change items from all scoring guide versions by score level.**

The change items in V1.6 were fairly well distributed among the remaining 4 scale degrees with approximately 20% each. The feedback was distributed a less uniformly with a higher emphasis at the top score level, 6, and the least emphasis at the new low score level, 2. The feedback also showed slightly more attention given to the titles and descriptions in the scoring guide than the change items.

*Evolution of the Scoring Guides.* By the end of 2010, there were three more draft versions of the scoring guides released beyond V1.6. The total numbers of unique changes for all draft versions of the revised scoring guides were charted in Figure 16. There is a clear difference in the number of change items in V1.6 and V1.8 compared to V1.7 and V1.9. Drafts V1.6 and V1.8 both followed the collection of detailed teacher feedback concerning the scoring

guides and these two draft versions have a much higher frequency of unique change items.

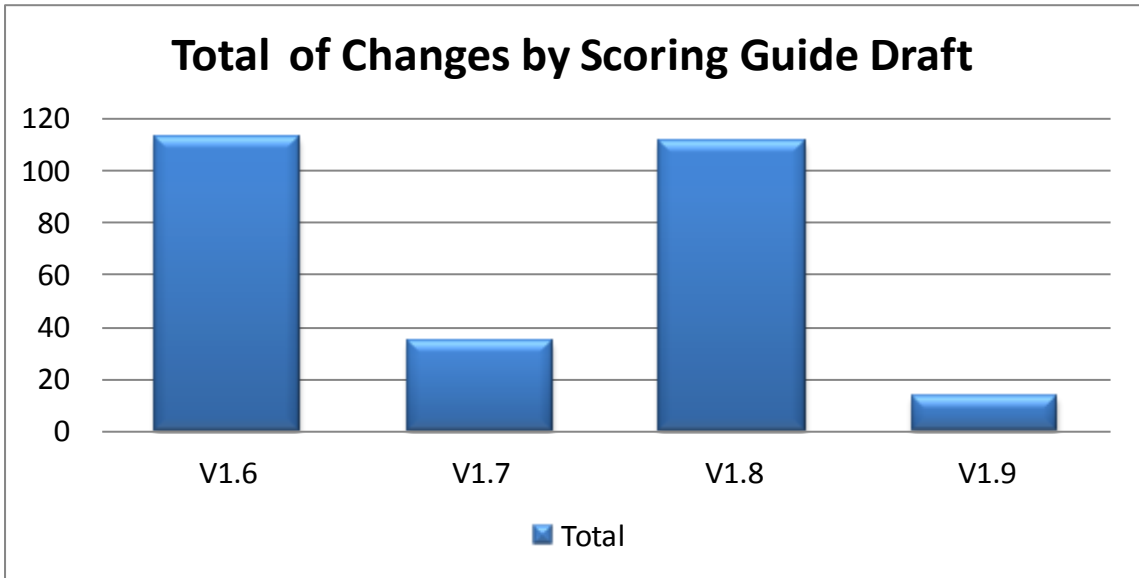


Figure 16: Total number of uniquely coded items by scoring guide draft.

The unique changes made to the versions of the scoring guides after V1.6 are shown in Figure 17.

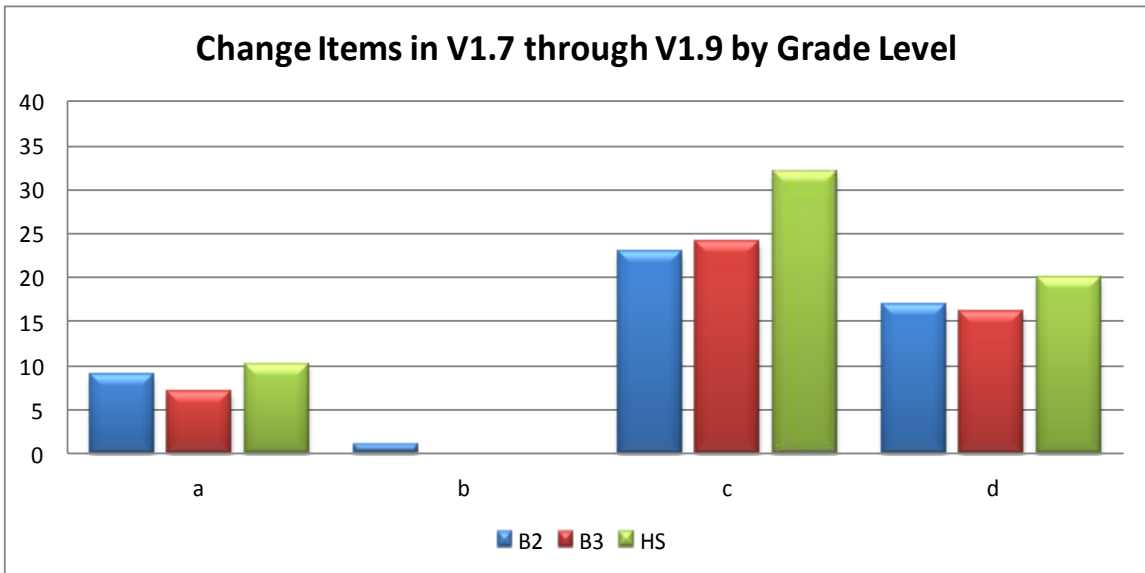
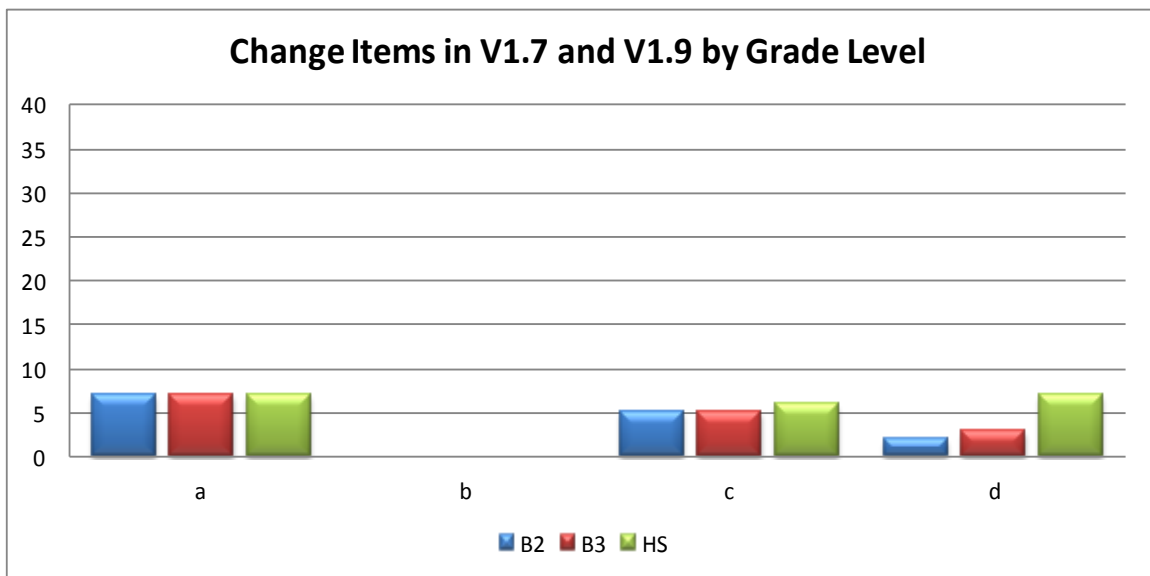
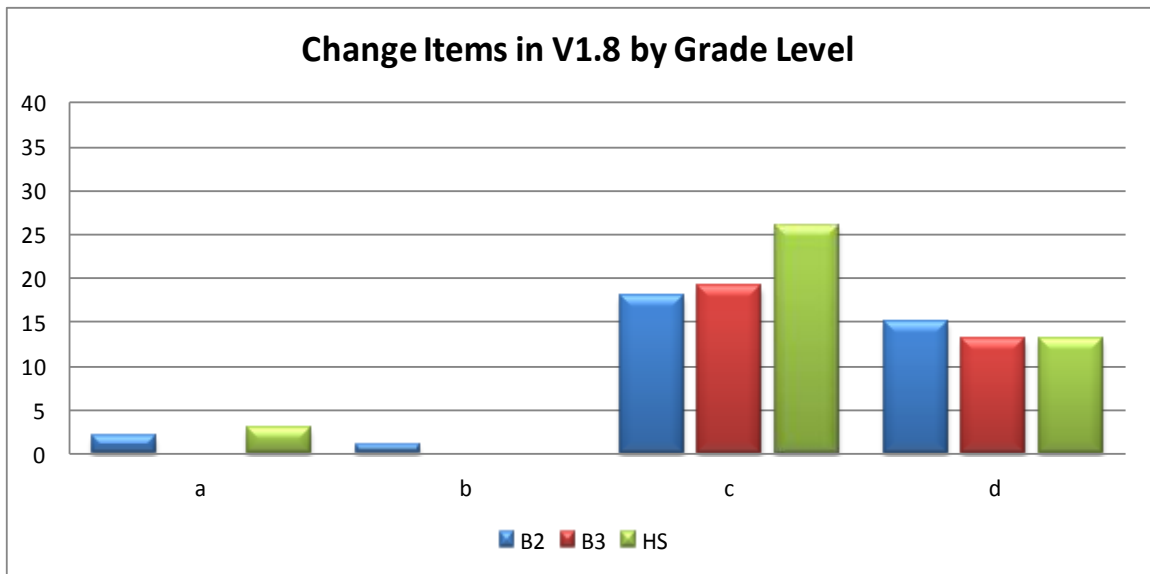


Figure 17: Unique change items in scoring guide drafts V1.7 through V1.9 by type.

However, since there was a second round of teacher feedback processed during the revision of draft V1.8, a more useful breakdown of this data is to look at V1.8, see Figure 19, separately from drafts V1.7 and V1.9. See Figure 18. Draft V1.9 showed a striking similarity to draft V1.7 in both quantity and types of unique change items. However, there was a dramatically different pattern in both the quantity and types of changes made to scoring guide drafts V1.7 and V1.9 compared to V1.6. In both V1.7 and V1.9 the number of change items was considerably fewer than V1.6. The changes that were identified were far more often to be of the grammatical type change, 'a1', than the feedback that was offered by the Teacher Evaluation Panels or changes made in V1.6.



**Figure 18: Unique change items in scoring guide drafts V1.7 and V1.9 by category and panel.**



**Figure 19: Unique change items in scoring guide draft V1.8 by category and panel.**

The changes identified in draft V1.8 had many similarities with the changes identified in V1.6. See Figure 5. In both drafts V1.6 and V1.8 the number of unique change items was considerably greater than unique change items identified in V1.7 and V1.9.

There was only one structural change made during this period of the scoring guide development. This change was in V1.8, following feedback from the summertime PD meetings, and was the only instance in this study of an item coded 'b3', Structural - Order of Bullet Points.

Another clear contrast between V1.8 and the set V1.7 and V1.9 can be observed in Figure 20 and Figure 21. The types of change items identified in V1.8 were very rarely grammatical, 'a1', and were very similar in type and distribution to the types of changes which were identified in V1.6 and the

feedback items from the Teacher Evaluation Panels. See Figure 5 and Figure 3 respectively.

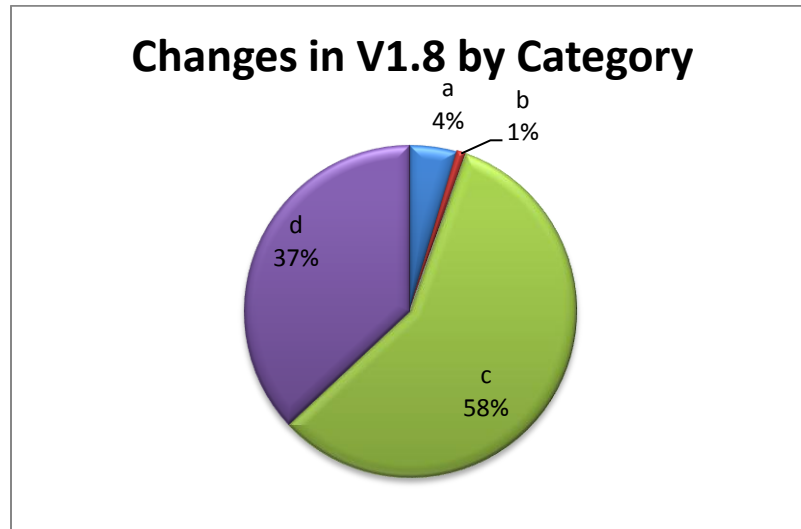


Figure 20: Changes in V1.8 by category.

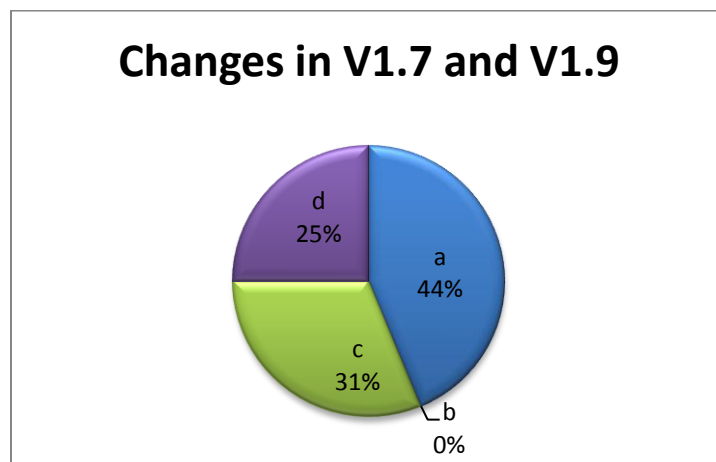


Figure 21: Changes to V1.7 and V1.9 by category.

The changes observed in V1.7 and V1.9 still show some functional type changes, categories 'c' - Evidence and 'd' - Performance Objective, were being



made during these revisions to the scoring guides. However, there is a clear contrast between V1.7 and V1.9 compared to V1.8 in both the proportionality of the changes in the 'c' and 'd' categories as well as the quantity of changes made in general.

The similarity between V1.6 and V1.8 is not as striking as V1.6 to the feedback. However, the both the quantity of changes identified in these drafts of the scoring guide and the concentration on the foundational and functional categories of change, categories 'c' - Evidence and 'd' - Performance Objective shows that V1.8 and V1.6 are much more similar than V1.8 compared to either V1.7 or V1.9. The context for the revisions of the scoring guides resulting in V1.6 and V1.8 was the receipt and inclusion of feedback from in-service teachers during the revision process for these drafts.

## Discussion

This case study set out to answer the following questions:

1. What types of feedback did the panels of experienced teachers offer ODE for the revision of the SI/ED scoring guides?
2. How did ODE utilize the feedback the teachers offered to revise the SI/ED Scoring Guides?

### *Conclusions:*

*Answering the First Question.* Through a process of careful text analysis, a Categorization Matrix was developed containing 12 codes defining different types of feedback items, and change items that were observed during this case study. This Categorization Matrix was used to code each item of feedback that the Teacher Evaluation Panels offered to ODE. Once all the feedback items were coded, several analyses were conducted to look deeper into the feedback and draw conclusions about the types of feedback the Teacher Evaluation Panels offered ODE.

The results showed the teacher evaluation panels feedback focused primarily on what the scoring guides were intended to measure. The most frequent feedback items were fairly evenly distributed among two categories of types: 1) clarifying or modifying the objective the students would be expected to perform and 2) clarifying or modifying the amount of evidence needed to determine the students' proficiency levels achieving the performance objective. Feedback of these types made up 80% of the total feedback offered by the Teacher Evaluation Panels. The remaining feedback was split evenly at 10% each for feedback items related to structural changes to the scoring guides or other items such as grammatical fixes. That is, the teachers' feedback primarily

addressed evidence for proficiency scores and the performance objectives rather than grammatical or other ancillary issues.

There were some clear differences between the feedback items offered by some of the Teacher Evaluation Panels. The middle school panels offered the most feedback overall. The middle school ED panel was the only panel to offer some types of feedback. The elementary ED panel offered considerably fewer items of feedback compared to the other panels. However, despite these variations, the panels were generally consistent with the types of feedback they offered, focusing on the scoring functions and the objectives of the scoring guides, denoted as categories 'c' - Evidence and 'd' - Performance Objective in the Categorization Matrix. These types of items were noted in the literature to be the first types of ideas to be considered when developing scoring guides, or rubrics (Mertler, 2001; Reddy, 2011).

Reddy (2011) and Mertler (2001) agreed on several steps they recommended to develop a scoring guide. Both researchers included a 'last' step in their lists, which was to work reiteratively with the scoring guide by looping through the recommended steps multiple times until the author(s) of the scoring guide was satisfied with the results. The revisions observed and analyzed in this case study exemplify this reiterative process and was a key feature of the case under study herein.

The literature also showed agreement on the first several steps in scoring guide development. The first of which is to define the objectives the students would be expected to perform, and which the scoring guides would be used to

measure. Next, closely related in importance, is to define the criteria, or evidence, needed to differentiate student scores. Based on these steps recommended by Reddy (2011) and Mertler (2001), the feedback items offered by the Teacher Evaluation Panels were foundational types addressing the basic functionality and usage of the scoring guides.

Reddy (2011) and Mertler (2001) also agreed on the importance of determining to what scale degree the students' performances would be assessed. Only two panels of six, the middle school ED and high school ED panels, recommended changes to the scale degree of the scoring guides, namely changing the scale degree from a 6 point scoring guide to a 4 point scoring guide.

Looking at the types of the feedback offered, 83% of the 166 unique feedback items identified in this study were of the types that directly paralleled the first steps in a development cycle as recommended by Mertler (2001) and Reddy (2011). This implies that the feedback offered by the Teacher Evaluation Panels was, at minimum, of the high quality type. How the ODE Science Content and Assessment Panels used the teachers' feedback was the next question in this study.

*Answering the Second Question.* Changes in the scoring guide draft versions were identified through a line by line comparison of blocks of text of each scoring guide with the blocks of text in prior draft version of the scoring guide. Then the identified changes were coded using the same method as was used to analyze and identify the types of feedback offered by the Teacher

Evaluation Panels. The analysis of these changes compared to the feedback documents was conducted looking for evidence supporting or refuting the inclusion of the teachers' feedback in the revised scoring guides.

Changes between the SI/ED Scoring Guide drafts V1.5 and V1.6 were made after the ODE Science Content and Assessment Panels received the feedback documents from the Teacher Evaluation Panels. Following the identification of change items the next round of analysis was to compare the changes identified in V1.6 to the feedback offered by Teacher Evaluation Panels. An item by item comparison between these documents showed strong evidence for many items of feedback being incorporated into the revised scoring guide drafts V1.6. Some of the change items in V1.6 matched the feedback verbatim. Some changes contained the same idea as feedback but the wording was not always adopted. The most demonstrative and clearly observed changes to the scoring guides was the shift from a 6 scale degree to a 4 scale degree scoring guide was directly attributed to the Teacher Evaluation Panels' feedback during the interview with the ODE staff member. Further the ODE staff member made clear the ODE Science Content and Assessment Panels' high opinion of the quality of the teachers' feedback.

Of the 113 unique changes identified in V1.6, 53 of those closely matched the feedback offered by the Teacher Evaluation Panels concerning the same blocks of text. That computes to an acceptance rate of approximately 47% of the changes in draft V1.6 were directly related to feedback from Teacher Evaluation Panels.

There were more feedback items offered by the Teacher Evaluation Panels than change items identified in V1.6. Of the 166 unique feedback items, the overall percentage of those items that directly matched change items was approximately 32%. This rate of acceptance is strong with approximately 1/3rd of the feedback being adopted by ODE. However, looking at the panels individually, the middle school ED panel had the lowest matching rate at 10%. This was less than half the matching rate of the panel with the next lowest matching rate.

The low matching rate of the middle school ED panel was likely a result of this panel offering more radical feedback than the other panels. This panel was the only pane to offer feedback of the types 'Other - Redefined' (a3) which redefined the block of text and 'Structural - Number of Bullets' (b1) which would alter the shape of the scoring guide such that it would no longer have the same structure as the scoring guides at the other benchmark levels. If these more radical feedback items were thrown out from the comparison analysis as outliers, the overall percentage of accepted feedback for the middle school ED panel improved to 36% versus 32%. The change in acceptance rate only marginally improved. However, perhaps in particular, for a set of documents that have already been seasoned through several revision cycles, vis a vis V1.5, identifying these types of feedback as having a low likelihood of being incorporated in later revisions may offer some guidance on what instructions to give to an evaluation panel when making recommendations for changes to a scoring guide or similar document.

As the scoring guides continued to evolve throughout 2010, the analysis of the drafts subsequent to V1.6 added more evidence supporting that teacher feedback was utilized in the revision of the scoring guides and that the feedback had a substantial impact on the development of the scoring guides. The analysis of changes in scoring guide drafts V1.7 through V1.9 showed distinct patterns in the number of change items as well as the types of change items identified. The changes in drafts V1.7 and V1.9 were much fewer in number compared to v1.6 and V1.8. The types of changes in V1.7 and V1.9 were far more likely to be grammatical 'a1' type changes, which did not affect the meaning of the scoring guides, than 'c' - Evidence and 'd' - Performance Objective type changes, which did affect the student performance requirements. The opposite was found for both drafts V1.6 and V1.8 in which there were many more changes identified than in V1.7 and V1.9. Further those changes were far more likely to be 'c' or 'd' type changes addressing the student performance requirements, than 'a1' grammatical changes. When these results are considered along with the contextual events surrounding the draft revisions, there is strong corroborating evidence supporting the statements made by the principle ODE staff member who explicitly asserted that the teachers' feedback was used to revise both drafts V1.6 and V1.8.

*Implications.* One of the underlying questions implicit in the research questions for this study was whether or not there would be value added through the process of obtaining teacher feedback during the development of the scoring guides. There was strong support in the literature for the inclusion of teachers in

the process of developing assessment tools (Reddy, 2011, Day & Matthews, 2008). The best practices recommended by Reddy (2011) included 8 steps for scoring guide development; specifically steps 4 and 5 were to obtain the feedback of stake holders, such as end users of the scoring guide, and then to revise the scoring guide based on that feedback. Throughout the cases studied herein, ODE has performed two separate rounds of steps 4 and 5 recommended by Reddy (2011).

Further, the first steps in scoring guide development are to identify the learning (performance) objectives, identify the criteria describing the evidence needed for each level of performance, and define how many levels of performance would be differentiated for the criteria (Reddy, 2011, Mertler, 2001). The feedback collected from the teachers, as shown through the analysis of the Teacher Evaluation Panels' feedback and the analysis of changes made to scoring guide drafts V1.6 and V1.8, indicate that these three steps, or facets, of scoring guide development were the primary subjects of the teachers' feedback and the resulting changes to the scoring guides. This further supports a claim that the scoring guides were likely improved by the events of this study. Not only has feedback from stakeholders been collected and incorporated in the revision of the scoring guides, but the feedback that was collected was of the type that addressed the foundational elements of the recommended scoring guide development process. There was no formal analysis that could assert for sure that the scoring guides were improved, such as improved validity or reliability, or any other measure of improvement. However, the evidence and the analysis



conducted in this study suggest that because of the types of feedback offered, it is, at least, very possible that the scoring guides were improved by including teacher feedback in the development cycle.

A second question implicit in this study was how teachers' attitudes toward policy decisions and instruments, such as the Oregon SI/ED Scoring Guides which follow from policy decisions, might be affected by the inclusion or exclusion of teachers in the development process. The literature is clear about the adverse effects top down education reform can have on teachers. The adverse effects can include: a reduction in teachers' sense of wellbeing (Vallie & Buese, 2007), teachers' not being able to reconcile the demands of the reforms with their teaching practice and thereby simply not adopting the reforms (Lumpe et al., 2000), and the potential deterioration of the working culture leading to factions within the staff and high turnover within the school or district (Olsen & Sexton, 2009).

Within this study there was only anecdotal evidence supporting a positive affective response on the part of the teacher participants toward the opportunity to engage in this process of developing these scoring guides. Some of this evidence was captured in the audio recordings. For example, several participants made statements of appreciation for the opportunity to participate in these panels as a professional development opportunity or simply as a way to approach the new scoring guides in a group. Another participant asserted that there would have been significant teacher backlash had V1.5 of the scoring guides been issued without the feedback process. Though teacher backlash

from the new SI/ED Scoring Guides may still occur, this particular teacher seemed at least somewhat gratified to have teachers included in the process.

Another piece of anecdotal evidence supporting the benefit of these processes with respect to aiding implementation was a theme in the audio discussions of the teachers hypothetically applying the scoring guides to past or future student work samples. While this was not a theme particularly relevant to this study, it does imply a certain degree of buy-in on the part of the teachers regarding the potential utility of the scoring guides as a tool, and may be a good predictor that the scoring guides will be used by these teachers in the future.

Another feature in the data that became evident through analysis was the quantity and quality of changes did not show a specific trend toward a final product. In fact, after two rounds of feedback collection the quantity of changes observed in the subsequent draft versions of the scoring guides, V1.6 and V1.8, showed nearly identical number and types of changes. It is known that there was nearly twice the quantity of feedback items than change items from the first round of feedback with the Teacher Evaluation Panels. It is not known at this time how much feedback was collected from teachers in the second round of feedback collection that occurred during the summer time PD events. It is a reasonable speculation given the number of teachers involved in the summer time PD events, that there was a large quantity of feedback collected during the second round of feedback. It is also known that the teachers' feedback from the summer time PD events was used to make changes to the scoring guides, which resulted in V1.8.

These results do not indicate a clear progress toward a document that represented a consensus of view of the teachers and ODE as to what the SI/ED Scoring Guides should be. Rather the fact that the number of changes did not decrease after the first round of feedback integration more likely indicates that the teachers in the summer time PD events were still unsatisfied with the scoring guides to a similar degree as the teachers in the Teacher Evaluation Panels.

The types of changes that occurred in V1.6 and V1.8, following teacher feedback, were predominantly focused on the function of the scoring guides, categories 'c' - Evidence and 'd' - Performance Objective in the Categorization Matrix. This also indicates that the teachers were still recommending changes that would make the scoring guides more functional in their opinion. As evidenced by feedback being accepted by the ODE Science Content and Assessment Panels, ODE agreed with some of this feedback. In this regard the scoring guides were moving toward a document that more closely represented what the teachers wanted and would, perhaps, better be able to use.

### *Limitations*

There were several issues which arose during this study that could challenge the results.

*Timing of the Interview.* The interview conducted with the ODE staff member occurred several months after the events of immediate interest for the interview had passed. The Teacher Evaluation Panels were convened in April, 2010. Draft V1.6 was released in June, 2010. The best time to conduct the interview would have been shortly after V1.6 had been released. The

interviewee's memory of specific feedback, the process of selecting feedback for inclusion, and the process of including the Teacher Evaluation Panels' feedback would have been more immediate. However, the interview was conducted after the release of V1.8, which included feedback from the summertime PD events. The time lapse between the interview and the work done with the first round of feedback compounded with the then recent work done with the second round of feedback somewhat confounds the issue of which feedback the interviewee was recalling and discussing during the interview. Through clarifying questions, some of the discussion concerning the feedback was made distinct between these two rounds of feedback. However, because both rounds of feedback were being discussed during the same interview, the subject of some general impressions and other comments made during the interview were difficult to resolve between the Teacher Evaluation Panels' feedback and the feedback collected during the summer time PD events. As a result the interpretation of some of these comments was subjective to the researcher's memory and impression of the intended meaning of the ambiguous comments collected during the phone interview.

The delayed timing of the interview did provide some additional data that were very useful concerning the second round of feedback from the summer time PD events. These data were not targeted in the semi-structured interview, but these data were instrumental in providing context for the changes observed in drafts V1.7 through V1.9. However, it would have been better had there been

two separate interviews to collect these data in order to address the two rounds of feedback individually.

*The Categorization Matrix.* The Categorization Matrix was developed in situ and ad hoc within the study and followed an evolutionary process with multiple revisions before the matrix was finalized and then applied to the entire data set. There were multiple attempts at the Categorization Matrix and trial analyses that were abandoned once they were found to be inadequate to meaningfully identify and distinguish the types of feedback and changes observed in the scoring guides. While this development process was intended to improve the final analysis there were some factors that could have been improved upon.

The themes sought within in the feedback documents were not considered in isolation from the themes sought in the change items within the scoring guide drafts versions. Even though the feedback and the scoring guides bore many similarities to the feedback items both structurally and linguistically, it is unclear whether or not the final Categorization Matrix would have been significantly different if it were developed solely to codify feedback items. Had the Categorization Matrix evolved along a feedback-centric path, the answer to the first question this study set out to answer may have been very different.

However, the end goal of the Categorization Matrix was to develop a system of codes that would facilitate the comparison between the feedback items and the change items. It seemed to be a necessary compromise to consider both the feedback and the scoring guide changes in the theme analysis leading

to identifying the general types in the Categorization Matrix. In fact, had the Categorization Matrix been developed using only themes identified in the feedback documents, the Categorization Matrix may not have resulted in a set of types that could be applied to the scoring guide changes and thereby would have not provided data that was useful to answer the second question for this study concerning how ODE might use the teachers' feedback. Given the similarities between the feedback items and the change items, it is unlikely that the Categorization Matrix would have been substantially different for the two data sets. However, this working assumption was worth noting here.

*Interpreting the Categorization Matrix.* The first attempt to organize the themes in the feedback and the scoring guides was to generate a rubric that would both categorize and attempt to capture the degree or severity of feedback and changes, moving from mild to substantial. In the final Categorization Matrix the categories themselves seem at first glance to fit a hierarchical progression. Though this interpretation could still be rationalized, it was not the final purpose of the Categorization Matrix to assert that the categories or types within the matrix increased in importance or significance moving from an 'a' type change to a 'd' type change along the horizontal axis or from a '1' to a '3' on the vertical axis.

In lieu of a rubric, the Categorization Matrix was developed to simply facilitate the codification of the feedback types and the types of changes made to the scoring guides. While the matrix retained some elements that look like an analytic rubric, such as an apparently progressive schema, this is simply an

artifact of the development process. The location of any given type indicated within the matrix should be considered arbitrary in comparison to any other type indicated in the matrix. That is, the matrix was not intended to weigh one type of change as more or less substantial compared to any other type of change. That said, the final analysis and a return to the literature did in fact indicate that items coded from category 'c' - Evidence and 'd' - Performance Objective were in fact more meaningful when compared to the rubric design recommendations in the literature. During the coding process, the value of any one code was considered the same as any other.

*Non-Conforming Feedback:* Most of the feedback offered by the Teacher Evaluation Panels was in one way or another marked on the scoring guide templates. However, some panels offered additional comments at the bottom of the page, and in some cases separate documents. All of these feedback items of feedback were submitted to ODE. These additional notes and comments were of three possible types: 1) rationales for, or extended descriptions of, changes already detailed in the feedback marked on the scoring guide template. 2) New items of feedback, not marked on the template, but that could be correlated to locations on the template for which changes, if they occurred, could be predicted or identified post hoc and 3) auxiliary items that did not easily fit within the data structure.

In the first case above, these items were most often placed in the teacher notes column of the coding spreadsheet. In the second case above, the feedback was placed into the spreadsheets next to the cells that these items

were interpreted by the researcher to be addressing. For example the HS-ED panel offered this item of feedback in a side note: "Relevant - (from meets) redundant with "relate" in same statement". The feedback was clearly addressing the proficient score level (meets) for the problem definition subsection. This was an 'a1' type item for that block of text as it was seeking a way to eliminate wording that did not add meaning. The third type of auxiliary feedback was the most challenging to deal with as it didn't fit easily within the data structure. There was only one item like this and it was from the middle school ED panel. This item was the note to format the scoring guides so that they would fit on one page front and back. In addition to not fitting in the data structure, this particular item did not fit within the Categorization Matrix. This item was not included in the analysis.

It was difficult to capture the auxiliary items like these in the analysis spreadsheets. This was especially true for items that did not address any particular location within the scoring guides. There was undoubtedly more than one item of feedback within the auxiliary feedback text which were lost. The structured data system that was developed in this study to analyze the feedback and change items resulted in over 1200 lines of text which were successfully analyzed. This included almost all of the data that did not easily fit within the developed data structure.



*Recommendations:*

The results of this study suggest several areas of further research and study as well as some recommendations concerning methods for future research.

*Characterize Impact of Teacher Feedback.* This study answered questions concerning the type of feedback teachers would offer to a state department of education to modify a state-wide assessment tool and how that feedback was used. This study did not address the impact this feedback cycle had on the assessment tool. To answer the question of whether or not the scoring guides were measurably improved, more valid or more reliable, as a result of the teachers' feedback, a follow up study would be required. Research to achieve this end would be a study of the inter rater and intrarater reliability of the scoring guides in drafts V1.5 through V1.9 and then compared to the reliability of these drafts to observe any changes that may be attributable to teachers' feedback. Likewise a study and comparison of the validity of the scoring guide drafts could indicate that teachers' feedback affected the validity of the assessment tools. These studies, in tandem with the results of the current study, could support or refute a conclusion that the scoring guides were, in fact, improved by the teachers' feedback.

*Teachers' Attitudes.* Another area of research that would further inform the benefit of including teachers in the development of state wide assessment tools would be to investigate how the teacher participants' attitudes and usages of these scoring guides may have been affected by their participation in the

Teacher Evaluation Panels. While very little preliminary data was collected concerning the teachers' attitudes toward the reform initiatives represented by the scoring guides or their attitudes toward a state issued scoring guide in general. It could be useful however, to collect additional data from the 24 participants of this study to compare with teachers who were not included in the development processes described in this study in order to determine the extent to which the direct participation in these events contributed to their attitudes, positively or negatively.

Another related question would be to what degree negative attitudes toward top down education reform might be affected with the three possible cases of teacher involvement: 1) The teacher/subject was directly involved in the process. 2) The teacher/subject was not directly involved but was aware that teachers were involved in the process. 3) The teacher/subject was unaware that any teachers were involved in the process. These studies would make more clear how teachers' direct, or representative, involvement in policy initiatives might affect the successful implementation of policy initiatives to avoid negative issues of 'treat rigidity' (Olsen & Sexton, 2009). And also answer questions concerning to what degree, if any, teachers attitudes are changed toward their ability to teach (Lumpe et al., 2000) or their professional roles (Valli & Buese, 2007).

All the Teacher Evaluation Panels recommended assembling similar panels in order to conduct a vertical articulation of the study guides. Many of the teacher participants expressed an interest in returning to participate in these

vertical articulation panels in order to better understand the up and down stream expectations for the SI and ED process knowledge and skills assessed through the performance assessment and to smooth the transitions from benchmark to benchmark. Convening panels such as these would provide a clear opportunity to conduct these studies while simultaneously taking the teachers' opinions of work that remains to be done for the development of the scoring guides.

*Utility of the Categorization Matrix.* This case study became a case study by document analysis and the Categorization Matrix proved to be the central instrument used to conduct this research. Another study to build upon the work done in this study would be a validity and reliability study of the Categorization Matrix itself. The Categorization matrix was developed for this specific case study, however, additional research could answer questions concerning how generalizable this Categorization Matrix is for future studies of scoring guide development.

Within the Category Matrix, the content addressed by categories 'c' - Evidence and 'd' - Performance Objective and type 'b2' - Change Scale Degree match very closely with the best practices discussed by Reddy (2011) and Mertler (2001). Analyzing the feedback and changes to the scoring guides in this study using this Categorization Matrix allowed for the interpretation of some types of changes observed in the scoring guides as fundamentally addressing the function of the scoring guides and an implicit interpretation that other types of changes had a lesser impact on the development of the scoring guide. The Categorization Matrix as a tool could be used to rigorously study the

development of another set of scoring guides toward several ends including the impact of certain contributors, the productivity or effectiveness of a set of development iterations, or identifying the convergence, or non-convergence, towards consensus on a scoring guide.

### *Closing Remarks*

The evidence collected and analyzed in this study supports the conclusion that in-service teachers can have a substantial impact on the development of a state wide assessment tool. That is, there is strong evidence that, when provided an amenable state department of education and knowledgeable in-service teachers, there can be a highly successful feedback loop that can substantially impact, and arguably improve the end product. The evidence reported in this study is merely suggestive of the kind of impact the teachers' feedback may have had and leaves open many questions as to what end the feedback process may have been working, whether the scoring guides were measurably improved -- such as increased validity or reliability, and how teachers' attitudes toward the policy initiative represented by the SI/ED Scoring Guides might have changed as a result of teachers being involved in the development process.

More research is required to answer these questions. However, the results reported herein strongly suggest that research in these areas can be successfully conducted and can be, at least, perceived to be productive and beneficial to the stakeholder participants involved.

## References

- Anonymous. (2009). Classroom "Reality Check" Strengthens Draft Standards. *American Teacher*, 94(December).
- Baker, E. L., & F.O'Neil Jr., H. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education: Principles, Policy & Practice*, 1(1), 11.
- Barlow, C., "How One State Established School Library/Technology Standards", *School Library Monthly*, 2009
- Blank, R. K. (2005). *Surveys of Enacted Curriculum: Tools and Services to Assist Educators*. Council of Chief State School Officers, One Massachusetts Avenue, NW, Suite 700, Washington, DC 20001-1431. Tel: 202-408-5505; Fax: 202-408-8072; Web site: <http://www.ccsso.org>.
- Blank, R.K, Porter, A, & Smithson, J., (2001). "Results from Survey of Enacted Curriculum Project, Final Report". Council of Chief State School Officers.
- Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, 15(9), 1277–1288.  
doi:10.1177/1049732305276687
- Gallagher, C. (2000). A Seat at the Table: Teachers Reclaiming Assessment through Rethinking Accountability. *The Phi Delta Kappan*, 81(7), 502–507.
- Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.
- Kleckner, C. (2009). Science - Standards - Oregon Department of Education. Retrieved August 16, 2012, from <http://www.ode.state.or.us/search/page/?id=1577>
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher*, 20(8), 15-21.
- Liu, O., Lee, H-S., Hofstetter, C., & Linn, M. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, 13(1), 33–33–55.
- Lumpe, A. T., Haney, J. J., & Czerniak, C. M. (2000). Assessing Teachers'

Beliefs about Their Science Teaching Context. *Journal of Research in Science Teaching*, 37(3), 275–92.

Mertler, C. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*. Peer-reviewed electronic Journal. Retrieved March 20, 2010, from <http://pareonline.net/getvn.asp?v=7&n=25>

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, 23(2), 13-23.

National Research Council. (2011). *Framework for k-12 science education : practices, crosscutting concepts, and core ideas*. Washington: National Academies Press.

Phillips, L. (2009). *2009-10 State Scoring Guide Professional Development, Assessing Scientific Inquiry* [PowerPoint Slides]. Retrieved from

Olsen, B., & Sexton, D. (2009). Threat Rigidity, School Reform, and How Teachers View Their Work Inside Current Education Policy Contexts. *American Educational Research Journal*, 46(1), 9 –44. doi:10.3102/0002831208320573

Reddy, M. Y. (2011). Design and development of rubrics to improve assessment outcomes: A pilot study in a Master’s level business program in India. *Quality Assurance in Education*, 19(1), 84–104. doi:10.1108/09684881111107771

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance Assessments: Political Rhetoric and Measurement Reality. *Educational Researcher*, 21(4), 22–27. doi:10.2307/1177207

Svicarovich, J. & Kirk, J.(2008). “Proficiency-Based Instruction and Assessment, A Promising Path to Higher Achievement in Oregon Education”, Oregon Education Roundtable, white paper.

Test Administration Manual - Oregon Department of Education. (2009). Retrieved August 16, 2010, from <http://www.ode.state.or.us/search/page/?=486>

Thomas, G. (2011). A Typology for the Case Study in Social Science Following a Review of Definition, Discourse, and Structure. *Qualitative Inquiry*, 17(6), 511–521. doi:10.1177/1077800411409884

- Turner, M. (2008). Determining paths to success: Preparing students for experimental design questions on standardized tests. *American Biology Teacher*, 70(3), 140–152.
- Valli, L., & Buese, D. (2007). The Changing Roles of Teachers in an Era of High-Stakes Accountability. *American Educational Research Journal*, 44(3), 519–558. doi:10.3102/0002831207306859
- Vanderwall, K. (2009). *2009-10 Achievement Standards Summary – All Subjects*. Retrieved from <http://www.ode.state.or.us/search/results/?id=240>
- Vanderwall, K. (2011). *2009-2011 Science Assessment of the 2009 Science Content Standards*. Retrieved from <http://www.ode.state.or.us/search/results/?id=240>
- Wiggins, G., "Standards and Criteria", Educative Assessment, Josey-Bass Publishers, San Francisco, CA, 1998. (103-138)
- Thompson, S. (2001). The Authentic Standards Movement and Its Evil Twin. *Phi Delta Kappan*, 82(5), 358.

## Appendix A

### Category Matrix

	Other	Structure	Evidence for Proficiency Score	Performance Objective
	a	b	c	d
1	Grammatical change that does not affect the meaning of the statement.	Adds or deletes number of bullets describing a score.	Clarifies the degree of evidence for proficiency score.	Clarifies performance objective.
2	Category is unclear from text in document.	Changes scale degree of scoring system.	Omits degree of evidence for proficiency score, but does not change performance objective.	Omits requirement for performance objective.
3	redefined category/objective (incommensurate)	changes order of bullets within score	Adds degree of evidence for proficiency score, but does not change performance objective.	Adds requirement for performance objective.



## Appendix B

### Demographic Survey Instrument

**Name:** \_\_\_\_\_

## **ODE SI/ED Scoring Guide Teacher Survey**

Please take a few minutes to respond to the following survey questions. If you do not find an appropriate response for a question, please write one in.

**1) Please indicate your gender.**

- Female  Male

**2) Please indicate your ethnicity/race. (Indicate all that apply)**

- American Indian or Alaska Native  
 Asian  
 Black or African American  
 Hispanic or Latino/a  
 Native Hawaiian or Other Pacific Islander  
 White

**3) How many years have you taught science prior to this year?**

- Less than 1 year  
 1-2 years  
 3-5 years  
 6-8 years  
 9-11 years  
 12-15 years  
 More than 15 years

**4) What is the highest degree you hold?**

- Does not apply  
 BA or BS  
 MA or MS  
 Multiple MA or MS  
 Ph.D. or Ed.D.  
 Other (Specify)

**5) What was your major field of study for the bachelor's degree?**

- Elementary Education  
 Middle School Education  
 Science Education  
 Science

- Science Education and Science
- Other Disciplines (includes Education fields, Math, History, English, Foreign Languages, etc.)

**6) If applicable, what was your major field of study for the highest degree you hold beyond a bachelor's degree?**

- Elementary Education
- Middle School Education
- Science Education
- Science
- Science Education and Science
- Other Disciplines (includes Education fields, Math, History, English, Foreign Languages, etc.)

**7) What type(s) of state certification do you currently have? (Indicate all that apply)**

- Emergency, provisional or temporary
- Elementary/Early Childhood Certification
- Middle School Certification
- Secondary Certification, in a field other than science
- Secondary Science Certification
- National Board Certification

**8) Please briefly describe your current teaching assignment.**

---

---

---

---

**9) How long have you used scoring rubrics to score student science work?**

- Less than 1 year
- 1-2 years
- 3-5 years
- More than 15 years
- 6-8 years
- 9-11 years
- 12-15 years

**10) For your first science course of the day, how frequently do you use scoring rubrics to:**

	Never	Rarely	Often	Always
Communicate expectations to students	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assess student achievement to guide my instruction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assess student knowledge and skills following instruction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**11) For your first science course of the day, how frequently do you ask students to:**

	Never	Rarely	Often	Always
Make educated guesses, predictions, or hypotheses	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Define a problem and/or specify criteria for a solution	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Follow step-by-step directions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Collect data	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Change a variable in an experiment to test a hypothesis	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Organize information in tables or graphs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Analyze and interpret data	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Design their own investigation or experiment to answer a scientific question	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Design, build, and test an engineering solution	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Make observations or classifications	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Make a written report of results from a laboratory activity, investigation, experiment, or a research project	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Make a presentation of results from a laboratory activity, investigation, experiment, or a research project	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**12) Please briefly state and explain your opinion about using scoring rubrics to assess student science work.**

---

---

---

**13) Please use the following space to state any additional information you would like to include with this survey.**

---

---

---

---

Thank you!

## DRAFT1.4- 2010-2011 Scientific Inquiry/Engineering Design Scoring Guide Benchmark 3: Grades 6, 7 and 8

	<b>SI- Forming a Question or Hypothesis</b> <i>Based on observations and scientific principles, propose questions or hypotheses that can be examined through scientific investigation.</i>	<b>ED- Identifying and defining a Practical problem to be solved.</b> <i>Based on observations and scientific principles, formulate the statement of a practical problem that can be addressed through the process of engineering design.</i>
<b>6</b>	<ul style="list-style-type: none"> <li>• Forms a question or hypothesis that can be scientifically investigated and that generally points toward an understanding of scientific relationships.</li> <li>• Provides comprehensive background science principles and observations to establish a detailed context for this investigation.</li> <li>• The question or hypothesis clearly guides the design of an innovative investigation.</li> </ul>	<ul style="list-style-type: none"> <li>• Describes in detail a practical problem to be solved through the process of engineering design. The solution addresses a specific need identified through research.</li> <li>• Uses and applies relevant science principles to identify potentially viable solutions to the problem.</li> <li>• Explains given criteria within constraints or limits for a solution based on science principles, with supporting rationale.</li> </ul>
<b>5</b>	<ul style="list-style-type: none"> <li>• Forms a question or hypothesis that can be systematically examined through a scientific investigation.</li> <li>• Provides background science principles and observations to establish a context for this investigation.</li> <li>• The question or hypothesis clearly guides the design of an effective investigation.</li> </ul>	<ul style="list-style-type: none"> <li>• Describes in detail a practical problem to be solved through the process of engineering design, clearly tying constraints and science principles to the problem.</li> <li>• Uses relevant science principles to identify potentially viable solutions to the problem.</li> <li>• Explains given criteria within constraints or limits for a solution based on science principles.</li> </ul>
<b>4</b>	<ul style="list-style-type: none"> <li>• Proposes a question or hypothesis that can be examined through a scientific investigation.</li> <li>• Provides observations and/or scientific principles that relate to the question or hypothesis.</li> <li>• The question or hypothesis is specific enough to guide the design of an effective investigation.</li> </ul>	<ul style="list-style-type: none"> <li>• Describes a practical problem to be solved through the process of engineering design.</li> <li>• Describes relevant science principles that relate to the problem.</li> <li>• States given criteria and constraints to be applied to the solution.</li> </ul>
<b>3</b>	<ul style="list-style-type: none"> <li>• Proposes a question or hypothesis that cannot be adequately examined through a scientific investigation.</li> <li>• Provides limited observations and/or scientific principles that relate to the question or hypothesis.</li> <li>• The question or hypothesis lacks the clarity necessary to guide the design of the appropriate investigative procedures.</li> </ul>	<ul style="list-style-type: none"> <li>• Partially describes a practical problem to be solved through the process of engineering design.</li> <li>• Describes science principles that partially relate to the problem.</li> <li>• States given criteria and constraints for a solution in an overly general way.</li> </ul>
<b>2</b>	<ul style="list-style-type: none"> <li>• Proposes a question or hypothesis that cannot be examined through a scientific investigation.</li> <li>• Provides limited observations and/or scientific principles that partially relate to the question or hypothesis.</li> <li>• The question or hypothesis cannot guide the design of the appropriate investigative procedures.</li> </ul>	<ul style="list-style-type: none"> <li>• Describes a problem that is unable to be solved through the process of engineering design.</li> <li>• Describes science principles that do not relate to the problem.</li> <li>• States unrelated criteria and constraints for a solution.</li> </ul>
<b>1</b>	<ul style="list-style-type: none"> <li>• Does not propose a question or hypothesis.</li> <li>• Provides no or totally unrelated observations and/or scientific principles.</li> <li>• No investigation design can be inferred.</li> </ul>	<ul style="list-style-type: none"> <li>• Does not describe a problem.</li> <li>• Does not describe science principles.</li> <li>• States no criteria or constraints for a solution.</li> </ul>

**DRAFT- 2010-2011 Scientific Inquiry/Engineering Design Scoring Guide**  
**Benchmark 3: Grades 6, 7 and 8**

	<p align="center"><b>SI- Designing an Investigation</b></p> <p><i>Design a scientific investigation to gather data to respond to a question or hypothesis.</i></p>	<p align="center"><b>ED- Generate possible solutions</b></p> <p><i>Select an engineering solution from a range of possible options, and design an investigation to provide sufficient data to test the proposed solution.</i></p>
<b>6</b>	<ul style="list-style-type: none"> <li>• Proposes scientifically logical, safe, and ethical procedures in a precise and efficient design that maximizes resources which contribute to the outcome.</li> <li>• Thoroughly identifies relevant variables and defines a systematic investigative process that is clearly defined and adaptable if necessary.</li> <li>• Presents a design that will provide data of exceptional quality and quantity to address the question or hypothesis and to investigate possible relationships.</li> <li>• Proposes scientifically logical, safe, and ethical procedures in a precise and efficient design.</li> <li>• Thoroughly identifies relevant variables and defines a systematic, investigative process that has clearly defined procedures.</li> <li>• Presents a design that will provide data of sufficient quality and quantity to address the question or hypothesis and to investigate possible relationships.</li> </ul>	<ul style="list-style-type: none"> <li>• Describes a variety of possible solutions that are distinctly different.</li> <li>• Uses the concept of trade-offs to compare and evaluate possible solutions in terms of criteria, constraints and priorities.</li> <li>• Selects and defends a solution for testing based on a comprehensive review of the design and performance criteria and constraints.</li> <li>• Describes multiple possible engineering solutions which may be similar to one another.</li> </ul>
<b>5</b>	<ul style="list-style-type: none"> <li>• Proposes a scientifically logical, safe, and ethical procedure that can be easily followed.</li> <li>• Identifies the independent and dependent variables and controls relevant to the procedure.</li> <li>• Designs a scientific investigation that uses appropriate tools and techniques to collect data relevant to the question or hypothesis.</li> </ul>	<ul style="list-style-type: none"> <li>• Uses the concept of trade-offs to evaluate possible solutions in terms of criteria, constraints and priorities.</li> <li>• Selects and defends a solution for testing based on the design and performance criteria and constraints.</li> </ul>
<b>4</b>	<ul style="list-style-type: none"> <li>• Proposes a scientifically logical, safe, and ethical procedure that includes significant scientific or logical errors.</li> <li>• Only partially identifies the independent and dependent variables and controls relevant to the procedure.</li> <li>• Designs a scientific investigation with insufficient tools and techniques to collect data relevant to the question or hypothesis.</li> </ul>	<ul style="list-style-type: none"> <li>• Proposes possible engineering solutions to the problem identified.</li> <li>• Evaluates the proposed solutions in terms of design and performance criteria, constraints, priorities, and/or trade-offs.</li> <li>• Selects a proposed solution for testing.</li> </ul>
<b>3</b>	<ul style="list-style-type: none"> <li>• Uses limited scientific knowledge or does not use logical, safe, or ethical procedures in the proposed design.</li> <li>• Independent and dependent variables and controls relevant to the procedure may be present, but are not identified.</li> <li>• Designs a scientific investigation lacking the necessary tools and techniques to collect data relevant to the question or hypothesis.</li> </ul>	<ul style="list-style-type: none"> <li>• Describes only one possible engineering solution.</li> <li>• Makes limited use of design and performance criteria, constraints, priorities, and/or trade-offs to evaluate the solution.</li> <li>• Presents solution for testing that partially relates to criteria and constraints.</li> </ul>
<b>2</b>	<ul style="list-style-type: none"> <li>• Uses minimal or incorrect scientific knowledge and unacceptable procedures in the proposed design.</li> <li>• Variables and/or controls are not present.</li> <li>• No tools or techniques are stated.</li> </ul>	<ul style="list-style-type: none"> <li>• Gives an incomplete description of a solution.</li> <li>• Incorrectly uses of the concept of trade-offs to evaluate possible solutions in terms of criteria and constraints.</li> <li>• Presents solution for testing with unrelated criteria.</li> </ul>
<b>1</b>		<ul style="list-style-type: none"> <li>• Does not describe a possible solution.</li> <li>• Does not complete any comparison or evaluations.</li> </ul>



**DRAFT- 2010-2011 Scientific Inquiry/Engineering Design Scoring Guide**  
**Benchmark 3: Grades 6, 7 and 8**

	<p align="center"><b>SI- Collecting and Presenting Data</b> <i>Collect, organize, and display sufficient data to support analysis.</i></p>	<p align="center"><b>ED – Testing Solution(s) and Collecting Data</b> <i>Test solution(s) by collecting, organizing, and displaying data to facilitate the analysis and interpretation of test results.</i></p>	
<b>6</b>	<ul style="list-style-type: none"> <li>• Rigorously follows the specified procedure, monitors data quality and utilizes the best available tools and techniques.</li> <li>• Carefully collects and records detailed and annotated data in a consistent and organized manner with the appropriate level of precision.</li> <li>• Displays data in a manner that highlights information and patterns and supports interpretation of relationships.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a functional solution that thoroughly addresses the criteria and constraints and is appropriate for testing. Design may incorporate modifications made during construction.</li> <li>• Collects detailed and complete data relevant to the criteria and constraints using effective and advanced techniques to test a solution.</li> <li>• Presents data that is complete and facilitates a thorough evaluation.</li> </ul>	<b>6</b>
<b>5</b>	<ul style="list-style-type: none"> <li>• Rigorously follows the specified procedure that utilizes appropriate tools and techniques.</li> <li>• Carefully collects and records detailed data in a consistent and organized manner with the appropriate level of precision.</li> <li>• Effectively displays data that facilitates thorough analysis and interpretation.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that thoroughly addresses the criteria and constraints and is appropriate for testing.</li> <li>• Collects detailed data relevant to the criteria and constraints using effective techniques to test a solution.</li> <li>• Presents data that is complete and facilitates a thorough evaluation.</li> </ul>	<b>5</b>
<b>4</b>	<ul style="list-style-type: none"> <li>• Follows the specified procedure that utilizes appropriate tools and techniques.</li> <li>• Collects and records data in a consistent and organized manner.</li> <li>• Displays data in a manner that supports analysis and interpretation.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that adequately addresses the criteria and constraints and is appropriate for testing.</li> <li>• Collects data relevant to the criteria and constraints using appropriate techniques to test a solution.</li> <li>• Presents data that is complete and facilitates evaluation.</li> </ul>	<b>4</b>
<b>3</b>	<ul style="list-style-type: none"> <li>• Partially follows the specified procedure that utilizes tools and techniques that work, but may not be the most effective or appropriate.</li> <li>• Collects and records some data but in an inconsistent or disorganized manner.</li> <li>• Displays data that are incomplete or disorganized.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that does not adequately address the criteria and constraints and/or can only be partially tested.</li> <li>• Collects data partially relevant to the criteria and constraints and/or used partially appropriate techniques to test a solution.</li> <li>• Presents data that is incomplete or does not facilitate evaluation of the solution.</li> </ul>	<b>3</b>
<b>2</b>	<ul style="list-style-type: none"> <li>• Inconsistently follows the specified procedure that utilizes inappropriate tools and techniques.</li> <li>• Collects and records irrelevant data in an inconsistent and disorganized manner.</li> <li>• Displays data that are inaccurate, incomplete and disorganized.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that does not address the criteria and constraints and cannot be tested.</li> <li>• Collects data that is not relevant to the criteria and constraints and does not use appropriate techniques to test a solution.</li> <li>• Presents data that is incorrect and does not facilitate evaluation of the solution.</li> </ul>	<b>2</b>
<b>1</b>	<ul style="list-style-type: none"> <li>• Does not follow the specified procedure and uses inappropriate tools and techniques.</li> <li>• Data randomly collected or missing.</li> <li>• Does not display data.</li> </ul>	<ul style="list-style-type: none"> <li>• Does not construct a solution.</li> <li>• Does not collect or present data.</li> </ul>	<b>1</b>

**DRAFT- 2010-2011 Scientific Inquiry/Engineering Design Scoring Guide**  
**Benchmark 3: Grades 6, 7 and 8**

	<b>SI- Analyzing and Interpreting Results</b>	<b>ED- Analyzing and Interpreting Results</b>
	<p><i>Summarize and analyze data including possible sources of error. Explain results and offer reasonable and accurate interpretations and implications.</i></p>	<p><i>Summarize and analyze test data, evaluating sources of error or bias, evaluate the proposed solution, describe any limitations or suggested design improvements, supported by test data and engineering principles.</i></p>
<b>6</b>	<ul style="list-style-type: none"> <li>Analyzes all the data and forms a comprehensive explanation that relates the experimental results to other scientific information.</li> <li>Clearly communicates the conclusions including sources, magnitude, and significance of error.</li> <li>Suggests revised or extended investigations based on analysis of results.</li> </ul>	<ul style="list-style-type: none"> <li>Thoroughly evaluates the tested solution and testing process referencing all design and performance criteria, constraints, priorities, and trade-offs.</li> <li>Thoroughly explains to what extent the solution addressed the criteria and constraints.</li> <li>Identifies and explains in detail possible design improvements using scientific and engineering principles.</li> </ul>
<b>5</b>	<ul style="list-style-type: none"> <li>Analyzes relevant data and forms an explanation that relates the experimental results to other scientific information.</li> <li>Clearly communicates the conclusions and discusses possible sources and effects of error.</li> <li>Suggests further investigations based on analysis of results.</li> </ul>	<ul style="list-style-type: none"> <li>Evaluates the tested solution in terms of all design and performance criteria, constraints, priorities, and trade-offs.</li> <li>Clearly describes to what extent the solution addressed the criteria and constraints.</li> <li>Identifies in detail possible design improvements.</li> </ul>
<b>4</b>	<ul style="list-style-type: none"> <li>Analyzes relevant data and constructs an evidence-based explanation of the results.</li> <li>Clearly communicates the conclusions including possible sources of error.</li> <li>Considers the results in light of the question or hypothesis and suggests possible revision(s) with justification.</li> </ul>	<ul style="list-style-type: none"> <li>Evaluates the tested solution in terms of design and performance criteria, constraints, and identifies priorities and trade-offs.</li> <li>Describes to what extent the solution addressed the criteria and constraints.</li> <li>Identifies possible design improvements.</li> </ul>
<b>3</b>	<ul style="list-style-type: none"> <li>Partially analyzes the data. Constructs an overly general explanation of the results of the investigation.</li> <li>Communicates conclusions in a general manner; stated sources of error are irrelevant.</li> <li>Minimally relates results to question or hypothesis. Suggests relevant revisions, but without justification.</li> </ul>	<ul style="list-style-type: none"> <li>Partially evaluates the tested solution in terms of design and performance criteria, constraints, and identifies some priorities and trade-offs.</li> <li>Incompletely describes to what extent the solution addressed the criteria and constraints.</li> <li>Identifies simplistic design improvements.</li> </ul>
<b>2</b>	<ul style="list-style-type: none"> <li>Inaccurately analyzes the data. Constructs a simplistic explanation of the results of the investigation.</li> <li>Incompletely communicates conclusions; lacking possible sources of error.</li> <li>Does not relate results to question or hypothesis. Suggested revisions are irrelevant to the investigation.</li> </ul>	<ul style="list-style-type: none"> <li>Inaccurately or incompletely evaluates the tested solution in limited terms of design and performance criteria, constraints, priorities, and/or trade-offs.</li> <li>Limited evidence provided as to what extent the solution addressed the criteria and constraints.</li> <li>Identifies irrelevant design improvements.</li> </ul>
<b>1</b>	<ul style="list-style-type: none"> <li>Fails to analyze the data. Offers no explanation of the results of the investigation.</li> <li>No conclusions or sources of error stated.</li> <li>No results discussed. No suggested revisions.</li> </ul>	<ul style="list-style-type: none"> <li>Provides no evaluation of the tested solution.</li> <li>Provides no evidence as to what extent the solution addressed the criteria and constraints.</li> <li>Does not identify design improvements.</li> </ul>



**Appendix D**  
**2010-11 Scientific Inquiry/Engineering Design Scoring Guide DRAFT1.6**  
**Grades 6, 7 and 8**

	<p style="text-align: center;"><b>SI - Forming a Question or Hypothesis</b>  <i>Based on observations and scientific principles, propose questions or hypotheses that can be examined through scientific investigation.</i></p> <ul style="list-style-type: none"> <li>Forms a question or hypothesis that can be scientifically investigated and demonstrates understanding of scientific relationships.</li> <li>Provides background science principles and observations to establish a detailed context for this investigation.</li> <li>The question or hypothesis clearly guides the design of an effective and/or innovative investigation.</li> </ul>	<p style="text-align: center;"><b>ED- Identifying and defining a problem to be solved</b>  <i>Based on observations and scientific principles, formulate the statement of a practical problem that can be addressed through the process of engineering design.</i></p> <ul style="list-style-type: none"> <li>Describes in detail a problem to be solved through the process of engineering design. The solution addresses a specific need identified through research.</li> <li>Uses and applies relevant background information and science principles to identify potentially viable solutions to the problem.</li> <li>Explains criteria and constraints or limits to be applied to a solution based on science principles, with supporting rationale.</li> </ul>	<b>5/6**</b>
<b>4</b>	<ul style="list-style-type: none"> <li>Proposes a question or hypothesis that can be scientifically investigated.</li> <li>Provides observations and/or scientific principles related to the question or hypothesis.</li> <li>The question or hypothesis is specific enough to guide the design of an effective investigation.</li> </ul>	<ul style="list-style-type: none"> <li>Defines a problem to be solved through the process of engineering design.</li> <li>Describes background information and relevant science principles that relate to the problem.</li> <li>Identifies criteria and constraints to be applied to the solution.</li> </ul>	<b>4</b>
<b>3</b>	<ul style="list-style-type: none"> <li>Proposes a question or hypothesis that is incomplete but could be scientifically investigated.</li> <li>Provides background observations and/or scientific principles that partially relate to the question or hypothesis.</li> <li>The question or hypothesis lacks the clarity necessary to guide the design of an effective investigation.</li> <li>Proposes a question or hypothesis that cannot be scientifically investigated.</li> <li>Provides background observations and/or scientific principles that are not relevant to the question or hypothesis.</li> <li>The question or hypothesis cannot guide the design of an effective investigation.</li> </ul>	<ul style="list-style-type: none"> <li>Partially defines a problem to be solved through the process of engineering design.</li> <li>Describes science principles that partially relate to the problem.</li> <li>Identifies given criteria and constraints to be applied to a solution in an overly general way.</li> <li>Defines a problem that is unable to be solved through the process of engineering design.</li> <li>Describes science principles that do not relate to the problem.</li> <li>Identifies unrelated criteria and constraints to be applied to a solution.</li> </ul>	<b>3</b>
<b>1/2*</b>	<ul style="list-style-type: none"> <li>Provides background observations and/or scientific principles that are not relevant to the question or hypothesis.</li> <li>The question or hypothesis cannot guide the design of an effective investigation.</li> </ul>	<ul style="list-style-type: none"> <li>Describes science principles that do not relate to the problem.</li> <li>Identifies unrelated criteria and constraints to be applied to a solution.</li> </ul>	<b>1/2*</b>

\*\*5 for preponderance (most) completed, 6 for all completed.  
\*2 for preponderance (most) completed, 1 for less completed.

**2010-11 Scientific Inquiry/Engineering Design Scoring Guide DRAFT1.6**  
**Grades 6, 7 and 8**

	<b>SI- Designing an Investigation</b> <i>Design a safe and ethical scientific investigation to gather data to respond to a question or hypothesis.</i>	<b>ED- Generate possible solutions</b> <i>Select an engineering solution from a range of possible options, and design an investigation to provide sufficient data to test the proposed solution.</i>	
<b>5/6**</b>	<ul style="list-style-type: none"> <li>Proposes scientifically logical, safe, and ethical procedures in a precise and efficient design that maximizes resources which contribute to the outcome.</li> <li>Thoroughly identifies relevant variables (including controls) and defines a systematic investigative process that is clearly defined and adaptable if necessary.</li> <li>Presents a design that will provide data of exceptional quality and quantity to address the question or hypothesis and to investigate possible relationships.</li> </ul>	<ul style="list-style-type: none"> <li>Describes a variety of possible solutions that are distinctly different.</li> <li>Uses the concept of trade-offs to compare and evaluate possible solutions in terms of criteria, constraints and priorities.</li> <li>Selects and defends a solution for testing based on a comprehensive review of the design and performance criteria and constraints.</li> </ul>	<b>5/6**</b>
<b>4</b>	<ul style="list-style-type: none"> <li>Proposes a scientifically logical, safe, and ethical procedure that can be easily followed; any missing steps could be inferred.</li> <li>Identifies the variables and controls relevant to the procedure.</li> <li>Designs a scientific investigation that uses appropriate resources/materials and techniques to collect data relevant to the question or hypothesis.</li> </ul>	<ul style="list-style-type: none"> <li>Proposes possible engineering solutions to the problem identified.</li> <li>Evaluates the proposed solutions in terms of design and performance criteria, constraints, priorities, and/or trade-offs.</li> <li>Selects and explains why a proposed solution was selected for testing.</li> </ul>	<b>4</b>
<b>3</b>	<ul style="list-style-type: none"> <li>Proposes a partially scientifically logical, safe, and ethical procedure that includes significant scientific errors.</li> <li>Only partially identifies the variables and controls relevant to the procedure.</li> <li>Designs a scientific investigation with insufficient resources/materials and techniques to collect data relevant to the question or hypothesis.</li> </ul>	<ul style="list-style-type: none"> <li>Describes only one possible engineering solution.</li> <li>Makes limited use of design and performance criteria, constraints, priorities, and/or trade-offs to evaluate the solution.</li> <li>Presents solution for testing that partially relates to criteria and constraints.</li> </ul>	<b>3</b>
<b>1/2 *</b>	<ul style="list-style-type: none"> <li>Proposes a procedure that is illogical and difficult to follow.</li> <li>Variables and controls relevant to the procedure may be present, but are not identified.</li> <li>Designs a scientific investigation lacking the necessary resources/materials and techniques to collect data relevant to the question or hypothesis.</li> </ul>	<ul style="list-style-type: none"> <li>Gives an incomplete description of a solution.</li> <li>Incorrectly uses of the concept of trade-offs to evaluate possible solutions in terms of criteria and constraints.</li> <li>Presents solution for testing with unrelated criteria.</li> </ul>	<b>1/2*</b>

## 2010-11 Scientific Inquiry/Engineering Design Scoring Guide DRAFT1.6

### Grades 6, 7 and 8

\* 2 for preponderance (most) completed, 1 for less completed.

	<p style="text-align: center;"><b>SI- Collecting and Presenting Data</b> <i>Collect, organize, and display sufficient data to support analysis.</i></p>	<p style="text-align: center;"><b>ED – Testing Solution(s) and Collecting Data</b> <i>Test solution(s) by collecting, organizing, and displaying data to facilitate the analysis and interpretation of test results.</i></p>	
<b>5/6**</b>	<ul style="list-style-type: none"> <li>• Rigorously follows the specified procedure that utilizes appropriate tools and techniques.</li> <li>• Carefully collects and records detailed and annotated data in a consistent and organized manner with the appropriate level of precision.</li> <li>• Displays data in a manner that highlights information and patterns and supports interpretation of relationships.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that thoroughly addresses the criteria and constraints and is appropriate for testing. Design may incorporate modifications made during construction.</li> <li>• Collects detailed and complete data relevant to the criteria and constraints using effective and/or advanced techniques to test a solution.</li> <li>• Presents data that is complete and facilitates a thorough evaluation.</li> </ul>	<b>5/6**</b>
<b>4</b>	<ul style="list-style-type: none"> <li>• Follows the specified procedure that utilizes appropriate tools and techniques.</li> <li>• Collects and records data in a consistent and organized manner.</li> <li>• Displays data in a manner that supports analysis and interpretation.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that adequately addresses the criteria and constraints and is appropriate for testing.</li> <li>• Collects data relevant to the criteria and constraints using appropriate techniques to test a solution.</li> <li>• Presents data that is complete and facilitates evaluation.</li> </ul>	<b>4</b>
<b>3</b>	<ul style="list-style-type: none"> <li>• Partially follows the specified procedure that utilizes tools and techniques but may not be the most effective or appropriate.</li> <li>• Collects and records data in an inconsistent or disorganized manner.</li> <li>• Displays data in a manner that is incomplete or disorganized.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that does not adequately address the criteria and constraints and/or can only be partially tested.</li> <li>• Collects data partially relevant to the criteria and constraints and/or used partially appropriate techniques to test a solution.</li> <li>• Presents data that is incomplete or does not facilitate evaluation of the solution.</li> </ul>	<b>3</b>
<b>1/2*</b>	<ul style="list-style-type: none"> <li>• Inconsistently follows the specified procedure that utilizes inappropriate tools and techniques.</li> <li>• Collects and records irrelevant data.</li> <li>• Displays inaccurate, incomplete and disorganized data.</li> </ul>	<ul style="list-style-type: none"> <li>• Constructs a solution that does not address the criteria and constraints and cannot be tested.</li> <li>• Collects data that is not relevant to the criteria and constraints and does not use appropriate techniques to test a solution.</li> <li>• Presents data that is incorrect and does not facilitate evaluation of the solution.</li> </ul>	<b>1/2*</b>



**2010-11 Scientific Inquiry/Engineering Design Scoring Guide DRAFT1.6**  
**Grades 6, 7 and 8**

	<p align="center"><b>SI- Analyzing and Interpreting Results</b></p> <p><i>Summarize and analyze data including possible sources of error. Explain results and offer reasonable and accurate interpretations and implications.</i></p>	<p align="center"><b>ED- Analyzing and Interpreting Results</b></p> <p><i>Summarize and analyze test data, evaluating sources of error or bias, evaluate the proposed solution, describe any limitations or suggested design improvements, supported by test data and engineering principles.</i></p>
<b>5/6**</b>	<ul style="list-style-type: none"> <li>• Analyzes relevant data and forms a comprehensive explanation (including patterns and trends) and it relates the results to other scientific information.</li> <li>• Clearly communicates the conclusions including sources, magnitude, and significance of error.</li> <li>• Suggests further investigations based on an analysis of results.</li> </ul>	<ul style="list-style-type: none"> <li>• Thoroughly evaluates the tested solution and testing process referencing design and performance criteria, constraints, priorities, and trade-offs.</li> <li>• Thoroughly explains to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies and explains in detail possible design improvements using scientific and engineering principles and trends in the data collected.</li> </ul>
<b>4</b>	<ul style="list-style-type: none"> <li>• Analyzes relevant data and constructs an evidence-based explanation of the results.</li> <li>• Clearly communicates the conclusions including possible sources of error.</li> <li>• Considers the results in light of the question or hypothesis and suggests possible revision(s) with justification.</li> </ul>	<ul style="list-style-type: none"> <li>• Evaluates the tested solution in terms of design and performance criteria, constraints, and identifies priorities and trade-offs.</li> <li>• Describes to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies possible design improvements.</li> </ul>
<b>3</b>	<ul style="list-style-type: none"> <li>• Partially analyzes the data. Constructs an overly general explanation of the results of the investigation.</li> <li>• Communicates conclusions in a general manner; stated sources of error are irrelevant or overly formulaic.</li> <li>• Minimally relates results to question or hypothesis. Suggests relevant revisions, but without justification.</li> <li>• Inaccurately analyzes the data. Constructs a simplistic explanation of the results of the investigation.</li> </ul>	<ul style="list-style-type: none"> <li>• Partially evaluates the tested solution in terms of design and performance criteria, constraints, and identifies some priorities and trade-offs.</li> <li>• Incompletely describes to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies simplistic design improvements.</li> </ul>
<b>1/2*</b>	<ul style="list-style-type: none"> <li>• Incompletely communicates conclusions; stated sources of error are missing or irrelevant.</li> <li>• Does not relate results to question or hypothesis. Suggested revisions are irrelevant to the investigation.</li> </ul>	<ul style="list-style-type: none"> <li>• Inaccurately or incompletely evaluates the tested solution in limited terms of design and performance criteria, constraints, priorities, and/or trade-offs.</li> <li>• Limited evidence provided as to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies irrelevant design improvements.</li> </ul>

## Appendix E

### DRAFT1.4- 2010-2011 Engineering Design Scoring Guide Benchmark 3: Grades 6, 7 and 8

	<b>Notes:</b>
<b>ED- Identifying and defining a Practical problem to be solved.</b>	
	<i>Based on observations and scientific principles, formulate the statement of a practical problem that can be addressed through the process of engineering design.</i>
<b>6</b>	<ul style="list-style-type: none"> <li>• Describes in detail a practical problem to be solved through the process of engineering design. The solution addresses a specific need identified through research.</li> <li>• Uses and applies relevant science principles to identify potentially viable solutions to the problem.</li> <li>• Explains given criteria within constraints or limits for a solution based on science principles, with supporting rationale.</li> </ul>
<b>5</b>	<ul style="list-style-type: none"> <li>• Describes in detail a practical problem to be solved through the process of engineering design, clearly tying constraints and science principles to the problem.</li> <li>• Uses relevant science principles to identify potentially viable solutions to the problem.</li> <li>• Explains given criteria within constraints or limits for a solution based on science principles.</li> </ul>
<b>4</b>	<ul style="list-style-type: none"> <li>• Describes a practical problem to be solved through the process of engineering design.</li> <li>• Describes relevant science principles that relate to the problem.</li> <li>• States given criteria and constraints to be applied to the solution.</li> </ul>
<b>3</b>	<ul style="list-style-type: none"> <li>• Partially describes a practical problem to be solved through the process of engineering design.</li> <li>• Describes science principles that partially relate to the problem.</li> <li>• States given criteria and constraints for a solution in an overly general way.</li> </ul>
<b>2</b>	<ul style="list-style-type: none"> <li>• Describes a problem that is unable to be solved through the process of engineering design.</li> <li>• Describes science principles that do not relate to the problem.</li> <li>• States unrelated criteria and constraints for a solution.</li> </ul>
<b>1</b>	<ul style="list-style-type: none"> <li>• Does not describe a problem.</li> <li>• Does not describe science principles.</li> <li>• States no criteria or constraints for a solution.</li> </ul>

**DRAFT - 2010-2011 Engineering Design Scoring Guide**  
**Benchmark 3: Grades 6, 7 and 8**

	<b>ED- Generate possible solutions</b>	<b>Notes:</b>
<b>6</b>	<p>Select an engineering solution from a range of possible options, and design an investigation to provide sufficient data to test the proposed solution.</p> <ul style="list-style-type: none"> <li>• Describes a variety of possible solutions that are distinctly different.</li> <li>• Uses the concept of trade-offs to compare and evaluate possible solutions in terms of criteria, constraints and priorities.</li> <li>• Selects and defends a solution for testing based on a comprehensive review of the design and performance criteria and constraints</li> </ul>	
<b>5</b>	<ul style="list-style-type: none"> <li>• Describes multiple possible engineering solutions which may be similar to one another.</li> <li>• Uses the concept of trade-offs to evaluate possible solutions in terms of criteria, constraints and priorities.</li> <li>• Selects and defends a solution for testing based on the design and performance criteria and constraints.</li> </ul>	
<b>4</b> meets	<ul style="list-style-type: none"> <li>• Proposes possible engineering solutions to the problem identified.</li> <li>• Evaluates the proposed solutions in terms of design and performance criteria, constraints, priorities, and/or trade-offs.</li> <li>• Selects a proposed solution for testing.</li> </ul>	
<b>3</b>	<ul style="list-style-type: none"> <li>• Describes only one possible engineering solution.</li> <li>• Makes limited use of design and performance criteria, constraints, priorities, and/or trade-offs to evaluate the solution.</li> <li>• Presents solution for testing that partially relates to criteria and constraints.</li> </ul>	
<b>2</b>	<ul style="list-style-type: none"> <li>• Gives an incomplete description of a solution.</li> <li>• Incorrectly uses the concept of trade-offs to evaluate possible solutions in terms of criteria and constraints.</li> <li>• Presents solution for testing with unrelated criteria.</li> </ul>	
<b>1</b>	<ul style="list-style-type: none"> <li>• Does not describe a possible solution.</li> <li>• Does not complete any comparison or evaluations.</li> </ul>	



**DRAFT - 2010-2011 Engineering Design Scoring Guide  
Benchmark 3: Grades 6, 7 and 8**

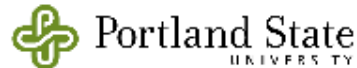
<b>ED – Testing Solution(s) and Collecting Data</b>		<b>Notes:</b>
	<i>Test solution(s) by collecting, organizing, and displaying data to facilitate the analysis and interpretation of test results.</i>	
<b>6</b>	<p>Constructs a functional solution that thoroughly addresses the criteria and constraints and is appropriate for testing. Design may incorporate modifications made during construction.</p> <ul style="list-style-type: none"> <li>• Collects detailed and complete data relevant to the criteria and constraints using effective and advanced techniques to test a solution.</li> <li>• Presents data that is complete and facilitates a thorough evaluation.</li> </ul>	
<b>5</b>	<p>Constructs a solution that thoroughly addresses the criteria and constraints and is appropriate for testing.</p> <ul style="list-style-type: none"> <li>• Collects detailed data relevant to the criteria and constraints using effective techniques to test a solution.</li> <li>• Presents data that is complete and facilitates a thorough evaluation.</li> </ul>	
<b>4</b> meets	<p>Constructs a solution that adequately addresses the criteria and constraints and is appropriate for testing.</p> <ul style="list-style-type: none"> <li>• Collects data relevant to the criteria and constraints using appropriate techniques to test a solution.</li> <li>• Presents data that is complete and facilitates evaluation.</li> </ul>	
<b>3</b>	<p>Constructs a solution that does not adequately address the criteria and constraints and/or can only be partially tested.</p> <ul style="list-style-type: none"> <li>• Collects data partially relevant to the criteria and constraints and/or used partially appropriate techniques to test a solution.</li> <li>• Presents data that is incomplete or does not facilitate evaluation of the solution.</li> </ul>	
<b>2</b>	<p>Constructs a solution that does not address the criteria and constraints and cannot be tested.</p> <ul style="list-style-type: none"> <li>• Collects data that is not relevant to the criteria and constraints and does not use appropriate techniques to test a solution.</li> <li>• Presents data that is incorrect and does not facilitate evaluation of the solution.</li> </ul>	
<b>1</b>	<ul style="list-style-type: none"> <li>• Does not construct a solution.</li> <li>• Does not collect or present data.</li> </ul>	

**DRAFT - 2010-2011 Engineering Design Scoring Guide  
Benchmark 3: Grades 6, 7 and 8**

	<b>ED - Analyzing and Interpreting Results</b>	<b>Notes:</b>
	<p><i>Summarize and analyze test data, evaluating sources of error or bias, evaluate the proposed solution, describe any limitations or suggested design improvements, supported by test data and engineering principles.</i></p>	
<b>6</b>	<ul style="list-style-type: none"> <li>• Thoroughly evaluates the tested solution and testing process referencing all design and performance criteria, constraints, priorities, and trade-offs.</li> <li>• Thoroughly explains to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies and explains in detail possible design improvements using scientific and engineering principles.</li> </ul>	
<b>5</b>	<ul style="list-style-type: none"> <li>• Evaluates the tested solution in terms of all design and performance criteria, constraints, priorities, and trade-offs.</li> <li>• Clearly describes to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies in detail possible design improvements.</li> </ul>	
<b>4</b> meets	<ul style="list-style-type: none"> <li>• Evaluates the tested solution in terms of design and performance criteria, constraints, and identifies priorities and trade-offs.</li> <li>• Describes to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies possible design improvements.</li> </ul>	
<b>3</b>	<ul style="list-style-type: none"> <li>• Partially evaluates the tested solution in terms of design and performance criteria, constraints, and identifies some priorities and trade-offs.</li> <li>• Incompletely describes to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies simplistic design improvements.</li> </ul>	
<b>2</b>	<ul style="list-style-type: none"> <li>• Inaccurately or incompletely evaluates the tested solution in limited terms of design and performance criteria, constraints, priorities, and/or trade-offs.</li> <li>• Limited evidence provided as to what extent the solution addressed the criteria and constraints.</li> <li>• Identifies irrelevant design improvements.</li> </ul>	
<b>1</b>	<ul style="list-style-type: none"> <li>• Provides no evaluation of the tested solution.</li> <li>• Provides no evidence as to what extent the solution addressed the criteria and constraints.</li> <li>• Does not identify design improvements.</li> </ul>	



## Appendix F Human Subjects Approval



Human Subjects Research Review Committee

Post Office Box 751  
Portland, Oregon 97207-0751

503-725-4288 tel  
503-725-3416 fax  
hsrrc@lists.pdx.edu

July 6, 2010

To: Emily Saxton

From: Nancy Koroloff, HSRRC Chair

Re: Approval of changes to your application titled, "An Investigation of the Reliability of the ODE Science Inquiry (SI) and Engineering Design (ED) Scoring Guide" (HSRRC Proposal # 101253).

Dear Emily,

In response to your request for an approval of change in your original HSRRC application, the Human Subjects Research Review Committee has reviewed your above-referenced project, last approved on April 16, 2010, for compliance with Department of Health and Human Services policies and regulations on the protection of human subjects. The committee is satisfied that your provisions for protecting the rights and welfare of all subjects participating in the research continue to satisfy federal requirements.

The following changes are approved: final ODE interview information (submitted 6/25/2010).

**Please be reminded that this project is due for continuing review two months before the expiration date of April 16, 2011.** Please submit a *Continuing Review Report* at that time (form is available in ORSP).

If you have questions or concerns, please contact the HSRRC in the Office of Research and Sponsored Projects (ORSP), (503) 725-4288, 6th Floor, Unitus Building, 4th & Lincoln.

Cc: William Becker

## Appendix G Semi-Structured Interview Guide

Interview conducted August 30, 2010.

The following questions concentrate on the development of the scoring guide's drafts V1.5 and V1.6 including the utilization or non-utilization of in-service teacher written recommendations based on V1.5 from the evaluation panels held April 2010.

1. Can you please describe your role in the development of the SI/ED scoring guides? Did you attend all the committee meetings.
2. How was teacher input utilized when drafting versions leading up to 1.5?
  - a. As I understand it, there were a few teachers in some of the draft committees, what role did those teachers, or other teachers, have leading up to draft 1.5?
3. The teachers in the April evaluation panels offered ODE structured feedback on version 1.5 of the SI and ED scoring guides. Can you characterize the process of how the feedback was utilized to revise the scoring guides?
4. The most noticeable change between drafts 1.5 and 1.6 was the transition from a 6 level scale to a 4 level scale. How influential was teacher feedback when making that change?
  - a. In particular, how influential was the feedback from the April evaluation panels in making this change?
5. What were the constraints or expectations ODE had for the development of the scoring guides?
6. What additional input or edits do you expect will come before the scoring guides are made official?
7. When is the board expected to vote to approve the scoring guides?
8. Looking back on the development of these documents, is there anything you would you do differently?
  - a. Was there anything that worked especially well and you would try to do again?
9. Are there any additional comments that you would like to offer about the development of the scoring guides or teacher input or feedback?

## Appendix H Feedback and Scoring Guide Coding Spreadsheet Example

SG	gra	cat	s	t	V.15	sc	notes	form	feedback	Teacher	sc	notes	V16	sc
SI	B3	question	6	1	SI: Forming a Question or Hypothesis	d1		quest or hyp --> Frame Investigat	SI: Framing the Investigation	Draft 15 was difficult to differentiate between 4/5/16.			SI: Forming a Question or Hypothesis	
SI	B3	question	6	1	Based on observations and scientific principles, propose questions or hypotheses that can be examined through scientific investigation.				<i>Based on observations and scientific principles, propose questions or hypotheses that can be examined through scientific investigation.</i>				Based on observations and scientific principles, propose questions or hypotheses that can be examined through scientific	
SI	B3	question	6	1	Forms a question or hypothesis that can be scientifically investigated and that generally points toward an understanding of scientific relationships.	d1	generally points to -> 'demonst rates'		Forms a question or hypothesis that can be scientifically investigated and demonstrates understanding of scientific relationships.	generally points to -> 'demonst rates'	d1		Forms a question or hypothesis that can be scientifically investigated and demonstrates understanding of scientific relationships.	100
SI	B3	question	6	2	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c3	adds appropriate		Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.				Provides background science principles and observations to establish a detailed context for this investigation.	
SI	B3	question	6	2	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c3	clarifies bgk		Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.				Provides background science principles and observations to establish a detailed context for this investigation.	
SI	B3	question	6	2	Provides comprehensive background science principles and observations to establish a detailed context for this investigation.	c2	omits comprehensive		Background research based on scientific principles and observations is appropriate and used to accurately establish a detailed context for the investigation.	Background is there for consistency and ease of	c2	omit comprehensive	Provides background science principles and observations to establish a detailed context for this investigation.	100

